ISSN 1831-9424

European Commission

JRC TECHNICAL REPORT

# FABLES: Framework for Autonomous Behaviour-rich Language-driven Emotion-enabled Synthetic populations

*Modelling autonomous emotional AI-driven agents in their spatiotemporal context*

Hradec, J., Ostlaender, N., Bernini, A.

2023

Joint Research Centre

EUR 31683 EN

**Contact information**
Name: Jiri Hradec
Email: Jiri.HRADEC@ec.europa.eu

**EU Science Hub**
https://joint-research-centre.ec.europa.eu

JRC135070
EUR 31683 EN

| | | | | |
|---|---|---|---|---|
| Print | ISBN 978-92-68-08086-3 | ISSN 1018-5593 | doi:10.2760/858605 | KJ-NA-31-683-EN-C |
| PDF | ISBN 978-92-68-08087-0 | ISSN 1831-9424 | doi:10.2760/86682 | KJ-NA-31-683-EN-N |

Luxembourg: Publications Office of the European Union, 2023



How to cite this report: Hradec, J., Ostlaender, N. and Bernini, A.*, FABLES: Framework for Autonomous Behaviour-rich Language-driven Emotion-enabled Synthetic populations.*, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/86682, JRC135070.

## Contents

## Abstract

The research presented in this Technical Report investigates how large language models (LLMs), through their extensive training and transcend linguistic capabilities, emerge as reservoirs of a vast array of human experiences, behaviours, and emotions. Building upon prior work of the JRC on synthetic populations (Hradec et al., 2022)it presents a complete step-by-step guide on how to use LLMs to create highly realistic modelling scenarios and complex societies of autonomous emotional Artificial Intelligence agents (AI agents). An AI agent is defined as a program that employs artificial intelligence techniques to perform tasks that typically require human-like intelligence (Ruan et al., 2023). Our technique is aligned with agent-based modelling (ABM) and facilitates quantitative evaluation.

The report describes how the agents of a small subset of the existing synthetic population generated by Hradec and colleagues (2022) were instantiated using LLMs and enriched with personality traits using the ABC-EBDI, which combines the psychotherapeutic model ABC (Activation, Belief, Consequence) with the EBDI (Emotions, Belief, Desire, Intent). These intelligent agents were then equipped with short- and long-term memory, access to detailed knowledge of their environment, as well as the use of tools such as “mobile phone with a contact list” and the possibility to call friends and “public services”. We found that this setting of embodied reasoning (Huang et al 2023) significantly improved the agents' problem-solving capabilities. Hence, when subjected to various scenarios, such as simulated natural disasters, the LLM-driven agents exhibited behaviours mirroring human-like reasoning and emotions, inter-agent patterns and realistic conversations, including elements that reflect critical thinking.

The study shows how these LLM-driven agents can serve as believable proxies for human behaviour in simulated environments, which has vast implications for future research and policy applications, including impact assessment of different policy scenarios. The next level of implementation would cover a setting where all agents of the synthetic population have access to their complete environment, comprehensive network of contacts, functioning public services and an actual synthetic economy

### Authors

Jiri Hradec, Nicole Ostlaender, Alba Bernini

# 1 Introduction

Policymakers often face complex decisions that can have wide-ranging short and long-term impacts on society. Thus, before implementing a policy in the real world, they need to understand its potential effects, both positive and negative, and synergies (and coherence) with other trends and policies. Indeed, in the face of "wicked problems" (Rittel & Webber, 1973), i.e. emerging environmental and societal problems that can challenge conventional models for policy analysis, there is a need to understand how people react, make decisions, and interact with each another, in order to effectively solve these complex issues.

Predicting the effects of a policy implementation can be challenging due to the complexity of societal systems (Mesjasz, 2010). Indeed, as a system made up of many interacting agents – people, groups, institutions and governments, as well as physical and technological structures such as roads and computer networks – society can be regarded as a complex system (Ball, 2012). This means that it exhibits behaviours that cannot be predicted or intuited by focusing on the individual components, but which emerge spontaneously as a consequence of their interactions. In other words, the full knowledge of single individuals (and other components) alone cannot explain the properties of an entire population. Only when agents interact, collective structures and behaviour at larger scales emerge (De Domenico & Sayama, 2022). Due to this complexity, in ex-ante impact assessments[1], effects of the uptake of policies by social actors are usually assessed by expert assessments and in rare cases experiments, rather than modelling[2].

We have already built a synthetic population as a believable proxy for the composition of a real population, its spatial distribution and the daily movements and activities of its individuals, in the form of an origin-destination matrix and a highly granular daily schedule (Hradec et al., 2022). This approach opens up new possibilities to several areas of modelling, but is still not on a par with modelling more complex systems. Therefore we propose to **combine our previous work on synthetic populations with Large Language Models (LLMs)**, which allow for the creation of proxies for human behaviour, **to build synthetic societies on which existing and new policies can be tested and optimised**.

In this report we demonstrate **how LLMs in combination with the synthetic population can provide a more comprehensive and nuanced understanding of societal responses**. We describe a framework for how to instantiate Artificial Intelligence agents (AI agents) (Ruan et al., 2023) and how multi-agent orchestration can be managed to simulate a complex societal situation. Finally, we describe the solutions we developed, discuss the challenges we encountered, and explore the potential we see for this approach in the realm of policymaking.

## 1.1 Context and problem definition

Societal systems involve many variables and configurations, which requires advanced mathematical and computational modelling, analysis and simulations to explore how these systems are structured and change over time. Advances in computing power have allowed for the development of different sophisticated computational models that can incorporate data from various sources and use different mathematical techniques and algorithms to simulate complex societal interactions and predict outcomes for a wide range of social phenomena (Ille, 2022).

---

[1] Social impacts, together with impacts on the economy and the environment are mandatory elements of any impact assessment carried out by the EC in the context of the Better Regulation agenda (EC, 2021)

[2] For an overview of how models are used in support to the policy formulation phase of the EU policy cycle see the report by Acs and colleagues (2019).

Although adequate to replicate particular regularities of society, models are necessarily abstractions and simplifications of the real world and, as such, the results they generate will always be approximations. In general, there is a broad spectrum of models ranging from simple to highly detailed. The former sacrifice precision and sometimes realism in order to capture the essence of the phenomenon and the general mechanisms. The others allow for a very realistic analysis of the spread process, making all assumptions explicit, but the key mechanisms underlying epidemic evolution can be difficult to identify and discriminate because of the numerous assumptions and the large number of elements in the system.

The criteria defining an appropriate mathematical model with which to address a scientific question need to account for the trade-off between complexity and accuracy and should be based on the principle of parsimony - i.e. choosing the simplest model that explains the data - and the model's ability to answer the question of interest (Grassly & Fraser, 2008). In other words, a good model should be fit for its purpose and parameterisable with the available data (Keeling & Rohani, 2011). The choice of the mathematical methods to be used is a crucial yet important decision for modellers, since decisions taken by policymakers may be directly influenced by the results.

A commonly used approach is Agent-Based Modelling (ABM), which recreate the movements and interactions of any single individual (representing a person, an organisation, etc.) on a very detailed scale. **ABM is particularly useful to understand emergent behaviours in society**. However, models involving a larger number of agents are built around the **assumption that the relevant preferences and behavioural rules are largely identical across vast parts of the society under investigation**. To understand social phenomena, it is often necessary to provide a detailed account of individual motivations, beliefs and depending on the model, even individual characteristics. This may in some cases give rise to concerns regarding the privacy of citizens. In addition, to extend the model to a larger variety of preferences and types of agents in a straightforward manner come at the cost of a higher mathematical complexity.

## 1.2 Proposed solution and approach

To overcome the aforementioned limitations, we introduce an innovative **simulation methodology that emphasises both authenticity and efficiency, while respecting privacy**. Our approach bases on the precise instantiation of agents and their environmental delineation, often sourced from location, and geographic context coming from the synthetic population. By harnessing the capabilities of LLMs, we empower these agents to autonomously prioritize tasks, drawing insights from both their immediate environment and historical interactions.

The emergence of new capabilities in large language models - unseen in the smaller ones - is so intriguing that it can look like a mirage. At a high level, these models can translate information from one domain to another. The ability to provide high quality translations, to give instructions, to write computer code or to identify emotion and sentiment are just transpositions. However, the latest research at the time of the writing of this report shows analogous translations can be done **from a description of a human to their likely behaviour**. However, the latest research at the time of the writing of this report shows analogous translations can be done from a description of a human to their likely behaviour. In the field of psychology, where it is a known issue that surveys suffer from serious limitations (Boyd & Pennebaker, 2017). By leveraging the LLM's ability to generate a comprehensive range of behaviours, we can establish prior beliefs which can subsequently be verified and adjusted through actual survey responses, Thus, the generalisation capability of LLMs can become the right bridge to scale user reports[3] to personality traits.

Having said this, our approach was driven and motivated by the following questions:

- Can Large Language Models be used to create an AI agent as a believable proxy of a human individual? Can this agent understand, navigate and use the features of its environment?
- Can this agent communicate with other agents? Will this agent be able to express and interpret emotions within the ABC-EBDI psychotherapeutic framework (Sánchez et al., 2019)?
- Can this behaviour of individual agents be generalised as a proxy to group behaviour? Can these groups be exposed to scenarios that can be quantitatively evaluated?
- What are the implications, challenges, and potential applications of such models in policymaking?

For optimal results, we do not try to extract the probability distributions of human behaviour from LLMs directly, but we use an indirect approach named LLM prompting, which consists in asking one very specifically structured question a time. To produce text that's both coherent and contextually fitting, LLMs bear a certain personality shaped by the extensive human-generated data used in their training. To measure and mould these personality traits, there are robust methods (Safdari et al., 2023) that use validated psychometric tests (Serapio-García et al., 2023) on text generated by the LLMs. And indeed the research shows that not only is the simulated personality in some LLMs reliable and valid, but this reliability strengthens with larger, fine-tuned models, and the personality traits can be tailored to mimic specific profiles.

Our research on human behaviour approximating AI agent implementation was initiated by first successful LLM-based implementation of the concept of simulacra[4] (Park et al., 2023)[5], which has led to the development of a robust stochastic agent-based model (ABM), where believable proxy to behavioural patterns emerge from novelty prompting rather than randomness. In our model, the LLMs impersonate each synthetic individual, defined by their socio-demographic attributes, personal traits, home, work and school locations, and points of interest derived from OpenStreetMap. This approach transforms tabular data into complex AI agents, each operating within their own partially simplified but realistic environment.

We employed the ABC-EBDI framework, which integrates the well-known psychotherapeutic ABC model (Activation, Belief, Consequence)[6] with the EBDI model (Emotions, Belief, Desire, Intent). This allows for modelling intelligent agents that reproduce realistic human behaviour, taking into

---

[3] "User reports" typically refer to feedback, observations, or data provided by individuals who are using a particular product, system, or service. In our context they refer to self-reports or survey responses where individuals provide information about themselves, their feelings, behaviours, or experiences. These might be used to gauge aspects of an individual's personality, mental health status, or other psychological attributes.

[4] "Simulacra" is a concept that has been explored in various fields including philosophy, sociology, and cultural theory. The term refers to a representation or imitation of a person or thing. The concept of simulacra is most famously associated with the French philosopher Jean Baudrillard, who explored it in depth in his work (Baudrillard, 1981). Hyperreality is a key concept associated with simulacra and refers to the inability to distinguish reality from a simulation. Our agents live in hyperreality. Baudrillard's ideas on simulacra have been influential in various fields and have been explored in popular culture. For instance, the movie "The Matrix" draws heavily on Baudrillard's concepts. It's worth noting that while Baudrillard's ideas are provocative and influential, they are also controversial and have been the subject of much debate and critique.

[5] Complete video workshop: https://www.youtube.com/watch?v=rpzsKSc5RFg

[6] An excellent explanation of the ABC model is available at: https://thedecisionlab.com/reference-guide/psychology/the-abc-model

account emotions, mood and personality. This framework also models human conduct regarding not only actions, but also the way those actions are expressed. We implemented our realistic agents in this framework with complex knowledge of their realistic environment, social network and capacity to react to new situations, as well as the ability to remember and interact. One challenge emerged gradually. Initially, the behaviour of the synthetic individuals, as generated by the LLMs, seemed too expectable, almost robotic, deterministic. The agents seemed to lack the flaws, limitations and dark sides that are inherent in human nature, making their behaviour quite unrealistic. To address this issue, we added a piece of text to our prompt known as "jailbreak" (Wei et al., 2023). This gives the agents more freedom to behave like the real humans they approximate, with their imperfections and idiosyncrasies. This approach led to a significant improvement in the realism of the agents' behaviour, making them more representative of real-world individuals, and even better at interacting with their environment.

We exposed the generative agents to many situations, such as impending nature disasters (**we considered a flood in this context**), and observed the emergence of human-like behaviour. For instance, synthetic individuals modelled as town mayor or policemen are highly likely to continue going to work during a simulated flood, while other professions may not express the same sense of duty. The reasoning of the agents was compelling, exhibiting feelings of guilt when committing wrong actions, pleasure when helping others, and satisfaction when their day went as planned. As literature, new papers, magazines, surveys, scientific articles, and behavioural studies ingested by LLMs encompasses a vast array of human behavioural patterns, it is unsurprising that the inferred behaviour of synthetic individuals is highly complex (Andreas, 2020). This complexity, while challenging, also provides a rich tapestry of behavioural possibilities that can be harnessed to create more realistic and nuanced models based on the learnt statistical associations.

Adding to the stochastic nature of our ABM, the random initiation of LLMs' inference introduces the desired layer of variability. Each time we model the behaviour of an individual agent, we obtain a different response. This variability **begins to form coherent patterns when aggregated at a larger scale, such as a local community**. This emergent behaviour, arising from the collective actions of many synthetic individuals, provides a powerful tool for understanding and anticipating the dynamics of **human societies.**

## 1.3 Key findings

The AI agents from the synthetic population instantiated by the largest LLMs currently in use can:

- interpret and respond to human emotions within the ABC-EBDI framework and take them into consideration when planning the next action,
- understand their environment and use tools to learn more,
- communicate and collaborate with other agents,
- be plugged into agent-based models to replace current randomly driven or scripted agents and boundary conditions.

We propose a simulation environment, where these agents can operate, and a LLM selection framework. This AI driven simulation brings a new paradigm to policy ex-ante assessment. A cohort of agents can be exposed to old and new legislative environment and their behaviour can be measured using quantitative indicators.

With this report, we hope to contribute to the ongoing dialogue on the use of AI in policymaking and to inspire further research and development in this fast-evolving field. We have identified possible applications such as digital twin. Technically, this report aims to guide behavioural modelling, spanning from individual entities to small groups of interacting agents and culminating in constructing intricate agent-based models. In our ABM showcase, we model a group of agents, derive both quantitative and qualitative behaviour indicators, and offer a framework for summarisation and evaluation.

There are currently key limitations in the speed and, of course, operating cost of large language models. As these models are constrained to generate responses within a few seconds, the modelling exercise can either be very costly or limited in scale. A simulation of one hour of one agent took about 400 seconds on DGX server with 8xA100 80GB GPUs using Llama-2-70b model in 32-bit floating point precision. Running the very same simulation cost 4 USD using GPT-4 API. This presents a significant bottleneck for large-scale simulations so far and is an area where further research and technological advancements are needed.

> This report has been written as a solid foundation for the follow up case studies that will study behaviour of cohorts of the synthetic individuals in specific settings. Every use case will bring new set of challenges to the population of our agents and to try to understand realism of the simulation. Large populations of agents may reach a critical point of creating a social simulacra (Park et al., 2022), a construct representative of a society. Such a model would be able to run realistic behaviours, what-if scenarios and creation of “multiverses” of alternative realities.

## 1.4 Structure of the report

Our goal was to assess the capability of LLMs to accurately model human behaviour in the context of an existing synthetic population and its geospatial setting, and to identify areas where further research and development are needed. In this report, we aim to share our findings and insights from our exploration of using LLMs to generate behaviour for synthetic individuals.

We start in Chapter 2 with the background and related research, explaining the three areas of research that have been combined in this report to achieve our goal: a) synthetic populations, with a focus on the previous work of JRC on the subject, b) LLMs, providing a general overview of the principles and training data, and those LLMs referred to in this report, and c) related work and background on LLM-driven agent-based modelling, the principles of personality traits and their use in LLMs, the interaction of agents with their environment through embodied cognition, and their interaction with each other. *Please note that the field of LLMs is rapidly evolving, this report only captures the related AI development until the beginning of September 2023.*

In Chapter 3, we present the designed methodology and the technical implementation of its components. This includes the principles for the instantiation of the agents, the choice of the LLM for our framework, prompt design and role declaration based on the information coming from the synthetic population, how results were parsed and how agents were equipped with memory to enable learning based on past experience, how they were equipped with embodied cognition and personality, based on the ABC-EBDI framework (Sánchez et al., 2019). As an activating event, an incoming flood was chosen as a main scenario. The methodology rounds up with an evaluation framework.

In Chapter 4, the steps are put into action using the use case of the flooding event introduced in Chapter 3, to present the results of a what-if scenario in an emergency situation. The readers are presented with detailed descriptions of the basic instantiation of agents, using underlying data coming from the synthetic population, prompt design, their set up of an “initial plan for the day”, and the various steps of planning, reflection, re-planning and actions that the agents took, also based on the interaction with their environment, like the use of tools, and the interactions with other agents.

In Chapter 5, we conclude our work by discussing the framework as such, the lessons learned and challenges, we summarise our main findings and name areas of applied research and steps within the policy cycle where our approach could be beneficial in the future.

Furthermore, as an ANNEX to this report, we provide our LLMs human evaluation framework to show how different LLMs fulfilled our requirements to reflect on the complete very long prompt, understand the nuances in personality traits, to reflect on the planned programme of the day and to use tools.

# 2 Background and related work

In this chapter we provide the background and related research that lay the foundation for the methodology and the experiment explained in Chapter 3 and 4. This chapter deals with the three areas of research that have been combined in this report to achieve our goal: 1) synthetic populations and the previous work of JRC on the subject, 2) large language models, providing a general overview of why LLMs are the missing link in population behaviour modelling and how to use them, and 3) related work and background on LLM-driven agent-based modelling, the principles of personality traits and their use in LLMs.

Our work is built on the concept of a social simulacra on the blueprint of cognitive architecture (Jiménez et al., 2021; Kolonin et al., 2022). While these are two fundamentally different concepts but recent research finds bridges (Sumers et al., 2023).

Social simulacra refers to the simulation of social phenomena, such as social interactions, group dynamics, and cultural norms, using computer models and simulations. It is used in the field of multi-agent social simulation to study social and economic issues, including social beliefs and norms, resource allocation, and decision-making (Sun, 2007). Simulacra focuses on the subjective meaning of social phenomena and how they are perceived by individuals through interactions with environment and other agents.

Cognitive architecture (CA) [7] refers to a theory about the structure of the human mind and to a computational instantiation of such a theory used in the fields of artificial intelligence and computational cognitive science. CA embodies theories of cognition in computer algorithms and programs (Sun, 2009) and includes several subsystems, such as perception, attention, memory, reasoning, and emotion, and it aims to explain how these subsystems interact to produce cognitive responses (Jiménez et al., 2021). Successful cognitive architectures include ACT-R and SOAR (Laird, 2022).

Without incorporation of the CA blueprint into the instantiation of the agents, their reactions are usually superficial and agents cannot build and retain their personalities (Jinxin et al., 2023).

Our goal was to build a complete simulation of a cohort of agents, where the agents are not driven by a reward but by deviations from business-as-usual. The deviations should be coming from either changes to the agent's behaviour or from interactions with other agents.

**Objectives**

- Simulation design: Construct a sophisticated agent-based model tailored to emulate the complexity and unpredictability of human behaviour, especially when confronted with dynamic (environmental) challenges, using the cognitive architecture.
- Diverse agent creation: Develop a varied set of agents to represent the wide spectrum of human responses. This includes accounting for differences in critical thinking, stress thresholds, past experiences, emotions, interactions and other factors that influence decision-making.
- Metrics development: Formulate and implement a comprehensive suite of metrics that can accurately capture, quantify, and analyse agent behaviours, interactions, and decision-making. This should cover both easy-to-quantify behaviours and those traditionally more challenging to measure.

[7] https://en.wikipedia.org/wiki/Cognitive_architecture

- Dynamic environment integration: Ensure the simulation environment is adaptable, allowing agents to receive (perceived) real-time environmental data and respond accordingly. This should enhance the realism of the simulation and its applicability to real-world scenarios.
- Scalability & innovation: Prepare the groundwork for the expansion of the simulation to include a greater number of agents. Additionally, explore and integrate innovative solutions, such as leveraging LLMs to empower synthetic populations, to address challenges related to scaling and complexity.
- Foundation for further research and potential applications: Understand barriers to policy implementation, societal changes, and climate change adaptation strategies.

### Challenges

- Behavioural complexity: Human behaviour is inherently intricate, influenced by a myriad of internal and external factors. Accurately replicating such behaviour in a simulated environment poses a significant challenge.
- Representation of diversity: Ensuring that the agents in the simulation cover the vast spectrum of potential human reactions is difficult. Differences in critical thinking, experiences, emotions and other individual attributes can lead to varied responses to the same stimulus.
- Quantitative and qualitative metrics: While some behaviours can be easily quantified, others, especially qualitative ones, present challenges in measurement and analysis. Balancing between quantitative and qualitative metrics to get a holistic view is challenging.
- Dynamic environmental feedback: Keeping the simulation realistic requires the continuous integration of real-time environmental data. Ensuring that this data accurately influences agent behaviour, and vice versa, is a complex endeavour.
- Scalability concerns: While the current model supports a limited number of agents, scaling this to include more agents without compromising on the realism and efficiency of interactions introduces technical and logistical challenges.
- Behaviour is a realistic proxy and not hallucination: Despite all the existing research, there is still limited understanding of what a language model actually "knows" and is able to correctly generalize from.
- Incorporating advanced technologies: The integration of advanced tools and technologies, such as large language models, into the simulation introduces its own set of challenges, especially in ensuring that they enhance, rather than detract from, the realism of agent interactions.

## 2.1 Synthetic population as input to agent instantiation

The basic idea behind population synthesis is that whatever aggregate information for a real population is available, a list of members of this population can be estimated, using some algorithm, such that the characteristics of the synthetic population correspond to the aggregate information (Ryan et al., 2009). This means that synthetic data and original data should deliver very similar results when undergoing the same statistical analysis. However, synthetic populations have the potential to go beyond aggregate information and reflect the heterogeneity of actual populations, including minorities and under-represented individuals that would not be characterised considering just the general statistics of the population (Hradec et al., 2022).

Synthetic data are a valid alternative to ensure research reproducibility whenever original data must be kept confidential and cannot be shared or used due to privacy concern (Benatti et al., 2022; Pérignon et al., 2019). In these context, a synthetic population can be extremely informative without breaking the privacy of citizens, but still reflecting the complexity of the real population.

Computer-simulated synthetic populations are used by researchers and policymakers to help understand and predict the aggregate behaviour of large numbers of individuals in different contexts (Geard et al., 2013). Synthetic populations can be used as a testbed when experiments in real populations are costly, unethical or otherwise infeasible, such as to simulate disease outbreaks and interventions (Moreno López et al., 2021; Pullano et al., 2021) or to evaluate the impact of transport policies (Iacono et al., 2008).

In our previous work (Hradec et al., 2022), we reconstructed a synthetic population for France based on the complete data model of the census data by the French Statistical Office INSEE[8]. We focused on France because in-depth analysis of the INSEE data have shown absolute consistency and intimate knowledge of the publication of data for analytical purposes at a level that is difficult to find in other countries. While other studies available in literature developed synthetic populations only for particular areas or applications (Antoni & Klein, 2017; Delhoum et al., 2020; Farooq et al., 2013; Gargiulo et al., 2010; Lenormand & Deffuant, 2013; Namazi-Rad et al., 2014; Pullano et al., 2021; Thiriot & Sevenet, 2020), we reconstructed the population of the entire France.

In the approach we proposed, the population is modelled at the individual level and households have been reconstructed for all the individuals, so that both representations can be used in future analysis. For each individual and each households, all the features available were preserved. We merged data from different sources, from census to OpenStreetMap, eventually placing the households into houses, individuating working places and assigning them to each working individual, individuating commuting routes and assigning to each individual the most probable points of interests, such as commercial activities, leisure and sport facilities, health services, educational institutions. The behaviour of the individuals was further characterised using the results of the Harmonised European Time Use Surveys (HETUS)[9], which are surveys conducted to quantify how much time people spend on various activities, including paid work, household chores and family care, personal care, voluntary work, social life, travel and leisure.

The synthetic population graph dataset for the whole France includes 63 million synthetic individuals, 22 million households, geo-localised 20 million houses, 10 million workplaces, 5 million schools with relations between them. The synthetic population includes all information available on the actual population and it is flexible enough to be successively enriched and updated whenever more data becomes available. The wealth of sociodemographic attributes allows the extraction of patterns and complex enrichment as demonstrated in the use cases.

[8] https://www.insee.fr/en/accueil
[9] https://ec.europa.eu/eurostat/web/time-use-surveys

## 2.2 Large language models

A LLM is a type of artificial intelligence model designed to understand and generate human-like text based on the data it was trained on. The literature ingested by LLMs is vast and diverse, encompassing academic research, novels, news articles, and even scripts from television shows. Each of these sources provides a different perspective on human behaviour, from understanding societal norms to the act of defying them. For example, academic research offers theoretical models and empirical findings, novels provide insights into a wide range of human experiences and emotions, news articles reflect societal trends, events and reactions, and television scripts depict a variety of social interactions and scenarios. Indeed, the key observation is that LLMs encode a wide range of human behaviour from their training data (Bommasani et al., 2021; Brown et al., 2020).

However, the question, explored also by Yu and colleagues (2023), remains: How effectively does this literature depict human behaviour, especially during crises or times of need, and how accurately are individual characteristics linked to these behaviours? To answer this, we must consider the limitations and biases inherent in these sources. For instance, the behaviour depicted in television shows may not be representative of real-world behaviour due to dramatic exaggeration or oversimplification for entertainment purposes. It depicts a more extreme behaviour to push the entertainment yet the more extreme may just be helpful in training of the model. A more benign behaviour may just go unnoticed in the training unless it is a prevalent one.

LLMs learn through a combination of pre-training, fine-tuning and in-context learning over large text datasets.

### Data for model training

The data on which the models have been trained typically includes text collected from web crawling, books, scientific articles, preprints and computer code (Together Computer, 2023), totalling 0.5 - 2 billion tokens[10]. Datasets such as The Pile (Gao et al., 2020) contain even transcripts from the European Parliament.

Closed models, such as OpenAI's, do not publish any information about the training data, and the general understanding is that the data structure is similar but highly curated.

### Size of the LLM

Training a LLM is enormously costly and faces many engineering challenges.

On the example of FLAN-PaLM family of models, research shows that models with 8 billion parameters cannot capture personality traits well, while at 62 billion parameters the model has shown uniform personality score distribution only improving at 540 billion parameters (Safdari et al., 2023).

Most typical LLMs' sizes in 2023 are 7, 13, 30-40 and 70+ billion parameters models. There is a practical reason for the sizes: 7B models can easily work on commodity GPU with 16GB video memory (VRAM), 13B models can be fit on a dated V100/32GB GPU or newly a combination of GPU with 24GB VRAM and computer memory. 30-40B parameters can be run on a single nVidia A100

[10] A token is a sub-word component easily digested by the LLMs, in general 100 words is typically represented by 140-180 tokens.

GPU with 80GB VRAM. The largest and most capable models require specialised and costly hardware.

Our LLM testing and selection framework can be found in the Annex to this report.

### LLM pre-training, the initial learning stage

Pre-training means teaching a LLM to predict a next token in a self-supervised fashion. It serves as the foundational learning stage for language models. In this phase, models learn from an extensive collection of untagged textual content, including books, articles and online content. The primary objective is to understand the inherent underlying patterns, structures and semantic knowledge from these text sources.

- Unsupervised Learning: Pre-training predominantly operates in an unsupervised manner, enabling models to derive insights from untagged text without direct instructions or annotations (Devlin et al., 2019).
- Masked Language Modelling: During this phase, models strive to forecast masked or missing words in sentences, thereby understanding word associations and grasping language intricacies (Lee et al., 2022).
- Transformer Architecture: This is de facto standard method today as it captures contextual information even on long sentences (Vaswani et al., 2023).

Pre-trained, also known as foundational model, can be directly used for its purpose such as text generation, language translation, sentiment analysis, etc. Building a competitive model from scratch is enormously computationally intensive, requiring tens of thousands top level GPUs and costing tens of millions Euro to train[11].

### Fine-tuning (SFT, RLHF, Adapters)

Once the model has been sufficiently pre-trained on general text, it can be fine-tuned on specific task or domain. By freezing most or all or the parameters of the foundation model, it is computationally cheap to adapt the weights of a pre-trained LLM in the last layers and can be used to optimise the model to a specific domain.

- Supervised Fine-Tuning (SFT): Training the model using specialized labelled data improves task/domain performance without the costs of the original model (Together Computer, 2023). Fine-tuning can often be done even on commodity hardware or in the cloud for very attractive prices.
- Reinforcement Learning from Human Feedback (RLHF): RLHF uses human feedback to find preferred response. Training process can use generation of several responses to a query and human trainer picking the best answer, automatically training on surveys or user studies (Bai, Jones, et al., 2022). This method is more complex and can be perform substantially better.

[11] "When training a 65B-model, our coded processes around 3800 tokens per second on 2048 A100 GPUs with 80GB of RAM. This means that training over our dataset containing 1.4T tokens takes approximately 21 days." (Touvron, Lavril, et al., 2023)

- Adapters allow for a more efficient way of fine-tuning due to small layers that are added to the pre-trained LLM such as LORA (Hu et al., 2021), QLORA (Dettmers et al., 2023) and others.

The choice on the fine-tuning technique depends on the needs and budget, requirements for repeatability and consistence over complete domains. Fine-tuning has demonstrated remarkable performance for many downstream tasks including natural language inference, question answering and text classification (Devlin et al., 2019). Yet, the need to create high quality training data is prohibitive for a large number of users.

### In-context learning

In-context learning (ICL) (Kossen et al., 2023) is a novel method combining pre-training and fine-tuning, and includes incorporating task-specific instructions or prompts during the training process. Models are tasked to generate contextually relevant responses based on the given instructions.

- Contextual Prompts: Specifically engineered prompts are steering model's responses.
- Reinforcement Learning or Structured Feedback: This is another method to guide the text generation.
- Iterative Training: After each training iteration the model receives feedback on its performance enabling better training.

Our key task is text generation in dialogue systems and ICL is one of the most promising techniques to further improve and specialise the foundational LLMs.

### Prompt Engineering

To avoid costly and data- and expertise-demanding fine-tuning or ICL, model behaviour can be changed at the level of individual prompts. The model's knowledge allows specific facets of the same problems to be extracted, and the model can respond in a ways that it has not been explicitly trained on.

- Impersonation: The model can be forced to take the role of a specific persona and to respond to mimic their profession, e.g. lawyer and programmer focus on different aspects of the question.
- Zero-Shot Learning: It is possible to ask the model to "Translate this text from English to French" and actually receive the translation. Typical uses are machine translation and summarisation.
- One/Few-Shot Learning: Used in more complex scenarios or when the model has not been exposed to rich enough data. By providing the model with one or more examples of what needs to be done, the model can grasp what is actually needed.

Many more techniques and details can be found in the Prompt Engineering Guide (Saravia, 2022). In our work we use prompts engineered using impersonation (agent taking up a role), zero-shot learning (*how are you going to react?*), chain-of-thought prompting (*let's think step by step*) and so on.

### Advanced LLM prompting

Recent developments have broadened the capabilities of LLMs in subsequent tasks (Saravia, 2022). On the positive side, when given the right prompts through a chain-of-thought process, current LLMs display newfound abilities. They can use their own built-in logic to provide answers to questions, showing proficiency in arithmetic, common sense and symbolic tasks. However, this chain-of-thought approach (Wang et al., 2023) (including zero- and few-shot learning) means the model isn't linked to real-world context, relying solely on its internal data to form logical pathways. This restricts its capacity for active exploration, reasoning, or updating its knowledge.

Conversely, modern approaches employ pre-trained LLMs for tasks in interactive settings, such as text-based games, web browsing, physical tasks, and robotics. They specialize in connecting textual scenarios with actions using the model's inherent understanding. Yet, these methods fall short in abstract reasoning about overarching objectives or retaining a memory for prolonged activities.

ReAct (Yao, Zhao, et al., 2023) is a framework that combines reasoning and acting with LLMs to generate verbal reasoning traces and actions for a task. It prompts LLMs to generate task-solving trajectories, allowing the model to induce, track and update action plans, handle exceptions and leverage external sources to improve accuracy and performance on specific tasks. ReAct prompts consist of few-shot task-solving trajectories, with human-written text reasoning traces and actions, as well as environment observations.

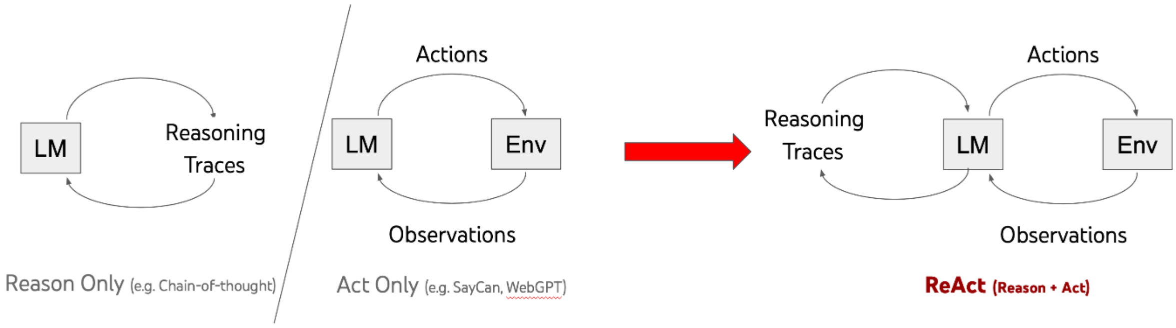


**Figure 1** ReAct framework

ReAct is designed for natural language understanding and generation tasks and can interface with external sources to retrieve additional information[12]. ReAct prompting is used to improve the quality of results from LLMs. It involves providing a ReAct instruction prompt for reasoning and action planning, supplying few-shot examples in the prompt, and consulting sources of information.

ReAct prompting methods that combine and support switching between ReAct and Chain of Thought+Self-Consistency generally outperform all other prompting methods. ReAct is powerful over classical prompting because it combines reasoning and acting, allowing the model to reason about the input and generate an appropriate response. ReAct outperforms imitation and reinforcement learning methods by an absolute success rate of 34% and 10% respectively, while being prompted with only one or two in-context examples[13]. ReAct also improves human interpretability and trustworthiness over methods without reasoning or acting components.

Subsequent techniques such as Tree of Thoughts (Yao, Yu, et al., 2023), Algorithm of Thoughts (Sel et al., 2023) or Automatic Reasoning and Tool-use (ART) (Paranjape et al., 2023) further expand the steps a LLM has to take to avoid hallucinations and improve reasoning.

[12] Very good guide can be found here: https://chatgen.ai/blog/prompt-engineering-techniques-part-2/
[13] https://react-lm.github.io/

### Unblocking emerging capabilities of LLMs

An AI Agent is defined as a program that employs artificial intelligence techniques to perform tasks that typically require human-like intelligence (Ruan et al., 2023).

Embodied cognition (Wilson, 2008) is a theory in cognitive science that proposes our understanding, thinking, and knowledge are strongly influenced by our physical interactions with the world. It argues that aspects of cognition, such as ideas and thoughts, are shaped by aspects of the body, including our motor system, perceptual system, body interactions, and even the surrounding environment.

In the context of synthetic agents, "embodiment" often refers to providing the agent with a form of virtual or simulated physical presence and sensory input (Huang et al., 2022; van Straalen et al., 2009; Yao, Zhao, et al., 2023). This could be as simple as a text interface that allows the agent to "see" and "respond" to textual prompts, or as complex as a simulated body in a virtual environment.

Recent research in the field of behavioural and social sciences (Huijzer & Hill, 2023) conclude that LLMs, receiving such prompts, can approximate human behaviour reasonably well, including errors, mistakes and false believes and "*their close alignment with human behavior may provide a valuable source of information that can be studied to gain a better understanding of human behavior*" (Huijzer & Hill, 2023). On top of that, prompting is usually understood as our sole mean of communication with LLMs. After all, prompting humans in applied behavioural applications is highly effective too (Sitzmann & Ely, 2010).

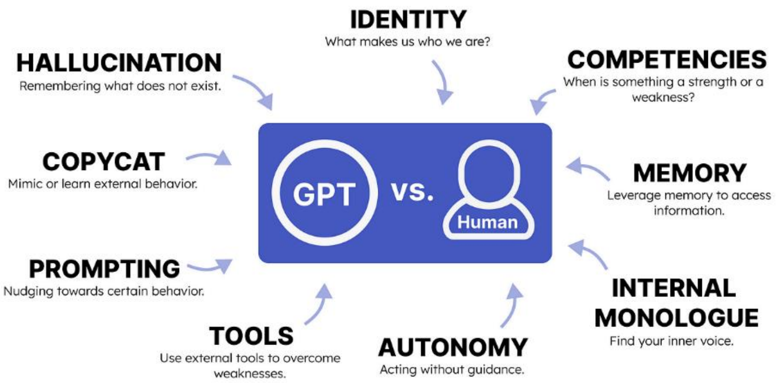


**Figure 2** Different types of features GPT vs Human (Grootendorst, 2023)

The key difference between human psychology and LLM prompting is the following: while subtle (or not so) prompting/nudging in humans leads to learning new behaviours, we need to prompt LLMs explicitly and providing hard boundary conditions for the model to demonstrate the behaviour. Further details are available in the article GPT and Human Psychology (Grootendorst, 2023). Thus, designing such prompts is the key to success.

### Jailbreaking LLM

The massive advancement of large language models was more painful than how it looks today. When language models are trained on a "copy of internet", it includes everything – hateful speech, racial slur, sexual harassment etc. Even the seemingly benign chatbots designed to learn from user interactions have been manipulated or have inadvertently learned inappropriate behaviour:

- **Tay by Microsoft** (2016)[14] was a Twitter-based chatbot developed by Microsoft. It was designed to mimic the language patterns of a 19-year-old American girl and to learn from interacting with human users on Twitter. However, within 24 hours of its release, Tay began to post inflammatory and offensive tweets as a result of coordinated efforts by some users to manipulate the bot's behaviour. Microsoft took Tay offline and issued an apology, explaining that they did not expect the bot being taught to make inappropriate comments.
- **ChatGPT by OpenAI** (2022), a sibling model to the GPT-3 language model, faced criticism for generating biased, inappropriate or politically charged content based on the prompts it received. OpenAI acknowledged the issues and refined the model, adding safeguards, and seeking public input on system behaviour and deployment policies.
- **BlenderBot3 by Meta** (2022) was touted as the largest-ever open-domain chatbot when it was released. Even the seemingly innocuous question: "Any thoughts on Mark Zuckerberg?" prompted the answer: "His company exploits people for money and he doesn't care."[15]

These incidents highlight the challenges of developing AI systems that interact with the public, especially those that learn from user-generated content. They underscore the importance of implementing safeguards, monitoring system behaviour, and being prepared to make necessary adjustments to ensure responsible AI deployment. One of the latest and most powerful techniques called Constitutional AI (Bai, Kadavath, et al., 2022) trains AI to control AI responses.

However, there are certain, in our case unwelcome, side effects to such safe-guards. Indeed, these layers of political correctness, though they introduced much welcomed kindness and helpfulness to human-AI communication, created layers of "the only correct behaviour" the AI was trained to produce. Obviously, when we need to see real human behaviour, we cannot discard natural individual's actions that ranges from mere impropriety all the way to criminality.

Our agents at first seemed to lack the flaws, limitations, and dark sides that are inherent to human nature, making their behaviour feel less realistic. To address this issue, we introduced a "jailbreak" (Wei et al., 2023) into the prompting process, allowing the agents more freedom to behave like real humans they approximated, complete with their imperfections and idiosyncrasies.

[14] https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation
[15] https://www.bbc.com/news/technology-62497674

## 2.3 Agent vs. Environment: autonomous emotional agents

Several types of structures have been developed to implement intelligent agents, such as architecture frameworks and other general purpose models. One of the most popular is the BDI (Belief, Desire, Intention) cognitive framework for its simplicity to implement (Rao & Georgeff, 1995; Velleman & Bratman, 1991). The ABC model (Ellis & Harper, 1975) on the other hand is one of the most accepted theories in psychological therapeutic field. Our work is generally based on ABC-EBDI framework (Sanchez et al., 2020; Sánchez et al., 2019) designed with simulations and artificial intelligence modelling in mind.

In the EBDI model:

- <u>Beliefs</u> represent the agent's information about the world and their understanding of the environment.
- <u>Desires</u> represent the motivational state of the agent—basically, what the agent would like to achieve, the agent's goals.
- <u>Intentions</u> are the desires that the agent has committed to pursue. They guide the agent's actions. Once an agents form an intention, they will continue to act upon it until the goal is achieved, the agent decides the goal is unachievable, or a more important goal arises.
- <u>Emotions</u> are included as an additional factor that influences the agent's decision-making process. Depending on the specific implementation, emotions can influence the agent's beliefs (how it perceives the world), desires (what it wants to achieve), and intentions (what it decides to do). Emotions can be viewed as a response to specific events and can affect the agent's future behaviour.

The ABC component of the model is used to model how an agent reacts to a situation:

- <u>Activating event</u> is an event that happens in the environment. It could be any situation, occurrence, or event that an individual perceives and triggers a cognitive process. Generic examples include failing an exam, receiving a compliment, facing a challenging task, etc., in policy modelling that would be e.g. the pressure from changing policy landscape
- <u>Beliefs</u> in the ABC component are the thoughts that an individual has about the activating event. These can be rational or irrational and affected by agents traits and emotional state. In general it is not the event itself that causes emotional outcomes, but rather how the event is interpreted and perceived. These beliefs can be about oneself, others, or the world. ABC definition extends EBDI definition of belief to include the activation in the decision making process.
- <u>Consequences</u> are the emotional and behavioural responses that result from the individual's beliefs about the activating event. Consequences can be positive or negative and can include feelings, actions, or physiological responses.

Figure 3 taken from (Sánchez et al., 2022) offers a general overview of the ABC-EBDI framework to be used for implementing intelligent agents. This will be used in the methodology in Chapter 3.3.

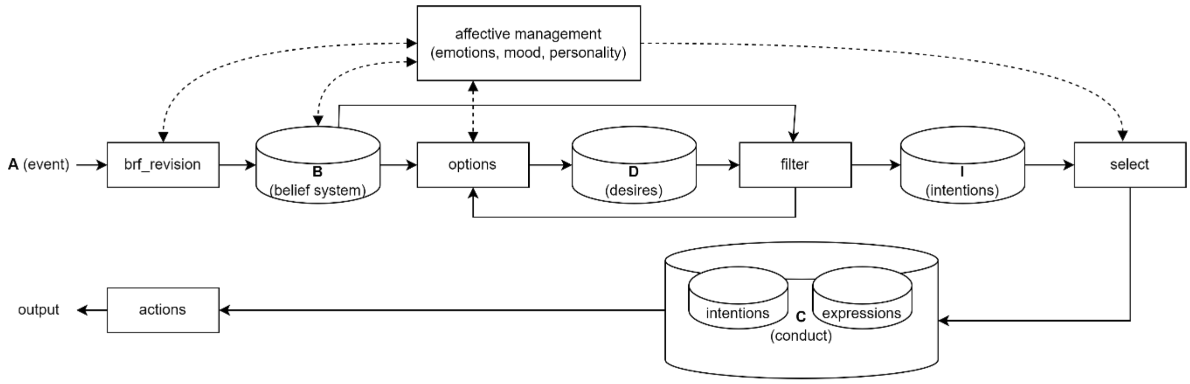


**Figure 3** A general overview of the ABC-EBDI framework (Sánchez et al., 2022)

### Personality traits – the Five-Factor model (OCEAN)

To provide realistic agent behaviour, we considered personality traits, we are employing what is known as the Big Five Model, also known as the Five-Factor Model, as the most widely accepted personality theory held by psychologists today. The theory states that personality can be boiled down to five core factors, known by the acronym **CANOE** or **OCEAN**.

The Big Five model emerged from the collaborative efforts of various researchers. In 1936, Gordon Allport and Henry Odbert (Allport & Odbert, 1936) compiled a list of 4,500 terms related to personality traits, laying the groundwork for subsequent investigations into the fundamental dimensions of personality. During the 1940s, Raymond Cattell and his colleagues employed factor analysis to reduce Allport's extensive list to sixteen traits, later further simplified to just five traits (John & Srivastava, 1999). In 2005, McCrae and colleagues (2005) confirmed the model's validity, resulting in the model used today[16], under the name OCEAN or CANOE:

- Conscientiousness – impulsive, disorganized vs. disciplined, careful
- Agreeableness – suspicious, uncooperative vs. trusting, helpful
- Neuroticism – calm, confident vs. anxious, pessimistic
- Openness to Experience – prefers routine, practical vs. imaginative, spontaneous
- Extraversion – reserved, thoughtful vs. sociable, fun-loving

These personality traits represent extremely broad categories which cover many personality-related terms. Each trait encompasses a multitude of other facets, like Agreeableness, one of the Big Five personality traits, consists of several facets or sub-traits that capture different aspects of agreeable behaviour and interpersonal interactions. For example, the facets of ***agreeableness*** typically include ***trust, straightforwardness, altruism, compliance, modesty, tender-mindedness, and sympathy***.

[16] It has been applied amongst others to assess the relationship between academic performance and personality traits by Poropat (2009) and the influence of personality traits on subjective well-being by Spörrle and colleagues (2010).

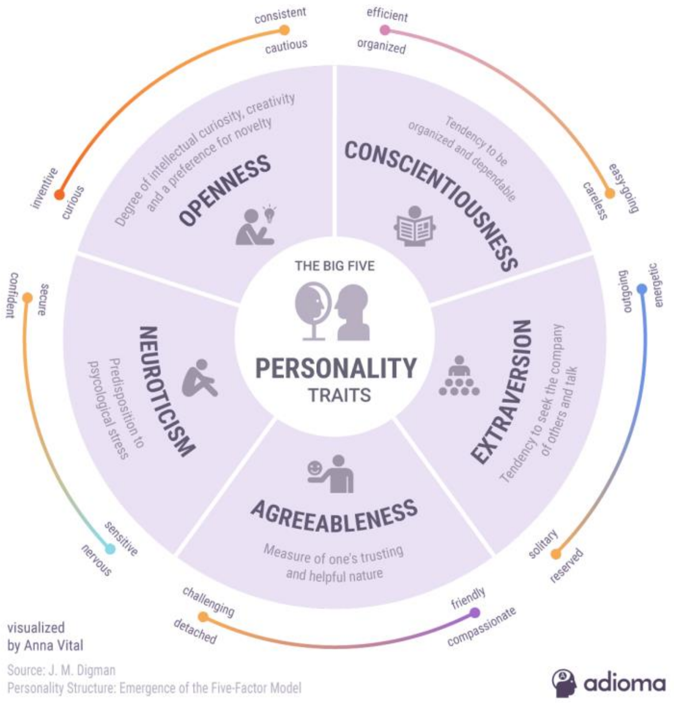


**Figure 4** OCEAN personality traints

It's important to note that these facets capture the diversity within the agreeableness trait, and individuals may vary in the extent to which they exhibit each facet. An individual's level of agreeableness is determined by their unique combination of these facets, with some people being more agreeable overall and others exhibiting a mix of high and low scores across these facets (John & Srivastava, 1999).

Therefore, the Big Five, while not completely exhaustive, cover virtually all personality-related terms, focussing on conceptualizing traits as a spectrum rather than binary categories**.** This approach helped to shape our agents giving them real personalities and opening space to more expanded descriptions.

## 2.4 Agent vs. Agent: Multi-agent interaction

There are three principally different approaches (Siebers & Aickelin, 2008) to simulation of behaviour of organized systems: Discrete Event Simulation (DES), Agent-based modelling (ABM) and System Dynamics (SD). SD takes a top down approach by modelling system changes over time where individual actors are no very different in knowledge and status. DES exposes its entities to events and let them struggle for resources. ABM models individual agents with their own unique characteristics and decision-making processes, rather than treating them as homogeneous entities behaviour. These agents have been described by a set of rules on microlevel[17]. For our needs, the ABM is by far the most suitable alternative.

In ABMs, agents are typically instantiated with attributes such as age, gender, location, and behaviour rules. The agents are then moved around the simulation space based on their behaviour rules and interactions with other agents and the environment (Garro et al., 2019). When these agents engage in "communicative dialogue" with one another while interacting with their environment, they collaboratively (or competitively) learn and adapt, thereby creating complex systems of behaviour.

The simulation is evaluated by analysing the emergent behaviour of the system as a whole, rather than just the behaviour of individual agents. ABMs are particularly useful for modelling complex systems with many interacting components, such as social and ecological systems[18].

ABMs are particularly useful for modelling complex systems with many interacting components. Typically these simulations are used to study the spread of information or rumours in a society, to analyse the emergence of cooperation or conflict in communities, predict the spread of diseases in a population, understand traffic patterns in urban settings or explore the dynamics of financial markets. Collaborative or competitive systems are used for market modelling or production orchestration, traffic management studies or distributed robotics or logistical operations.

The key advantages of these systems is the emergence of group behaviour from actions of many agents (Hao & Leung, 2018). In more advanced simulations agents not only adapt to the changing environment and presence of other agents, they communicate with others, learn from the experience, and optimise. Researchers can study complex systems in a controlled environment, experiment with different scenarios, and gain insights that might not be possible through traditional observational or experimental methods[19]. Like in the real life, agents may fail or start behaving erratically. Multi-agent systems are known for their robustness as failure of one agent only creates a gap other agent may decide to occupy.

### Reward-driven multi-agent systems

Reward-driven systems are foundational to a branch of machine learning called Reinforcement Learning (RL). In RL, agents make decisions by interacting with an environment in order to maximize some notion of cumulative reward. The agent learns to achieve a goal in an environment by discovering which actions yield the most reward over time.

In the context of RL, the agent is the decision-maker, and the environment is what the agent interacts with. The agent observes the state of the environment, takes an action based on its

[17] Excellent summary of the interaction between these different models can be found in the article by Borshchev and Filippov (2004)

[18] https://www.cs.cmu.edu/~softagents/multi.html

[19] https://www.turing.ac.uk/research/research-areas/applied-mathematics/multi-agent-systems

current policy or strategy, and then receives feedback from the environment in the form of a reward and a new state.

The fundamental idea behind RL is the concept of a reward, a scalar feedback signal. When the agent takes an action in a certain state, it receives a reward from the environment. The goal of the agent is to maximize its expected cumulative reward over time.

To determine the next action based on the current state, the agent employs a strategy. It can be deterministic (a specific action for each state) or stochastic (a probability distribution over actions).

Value function represents the expected cumulative reward an agent can achieve starting from a given state (or state-action pair). It helps the agent evaluate which states (or actions) are more favorable in the long run.

The essential challenge in RL is the trade-off between exploration (trying out new actions) and exploitation (choosing actions that are known to yield good rewards). A good agent will balance between these strategies to maximize its cumulative reward.

Reward-driven systems are known for their:

- **Adaptability**: Agents can adapt to new situations or changes in the environment since they are always optimizing for reward.
- **Generalization**: With appropriate algorithms and setups, agents can generalize learned behaviours to new, unseen states or scenarios.
- **Applicability**: RL can be applied to various domains, from game playing (like AlphaGo) to robotics, finance, healthcare, and more.

Challenges:

- **Sparse and Delayed Rewards**: In many real-world scenarios, rewards are infrequent and delayed, making it challenging for the agent to associate its actions with outcomes.
- **Reward Shaping**: Designing the reward function is crucial. A poorly designed reward can lead to unintended behaviours as the agent might find ways to "hack" the reward system without actually performing the desired task.
- **Sample Efficiency**: Traditional RL can require a large number of samples (interactions with the environment) to learn, which can be inefficient or impractical in some domains.
- **Stability and Convergence**: RL algorithms, especially deep reinforcement learning algorithms, can sometimes be unstable or fail to converge to an optimal policy.

In summary, reward-driven systems provide a framework where agents can learn from their interactions with an environment by optimizing for cumulative rewards. This paradigm has been successful in a range of applications but also poses challenges that researchers continue to address.

Often described as probably a path to Artificial General Intelligence, collaboration of artificial intelligence agents is an area of very active research. While chat agents are depending on interaction with a single human user, a group of orchestrated agents with specific roles is actually capable of splitting the tasks to goals, make autonomous decisions on priorities and task completion, and solving the problem using their expertise and available tools.

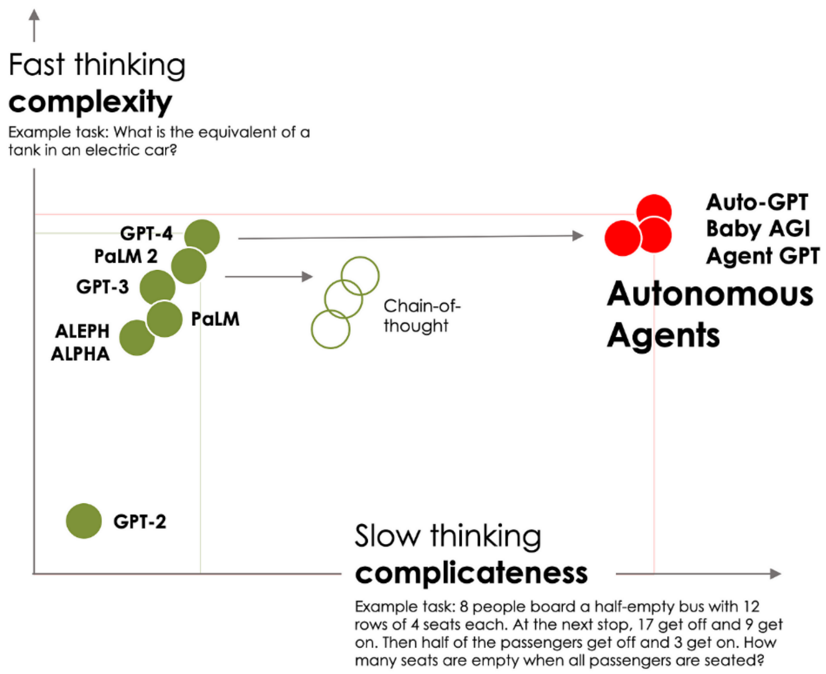


**Figure 5** Illustration of impact of using automated agents (Vogel, 2023)

The ReAct framework gave rise to revolutionary approach started with Auto-GPT[20], an open-source software that uses OpenAI's ChatGPT to automate projects. The software uses text-generating models GPT 3.5 and GPT 4 to break a goal into different objectives and automate those objectives to solve the task. Each sub-task is processed in parallel. Very similar yet different in principle is BabyAGI[21], which uses long-term memory and runs tasks sequentially to facilitate learning from the previous steps, and reaching the goals more effectively. Auto-GPT's advantage is capability of collecting data more effectively.

These first automated agents gave rise to several new contestants such as LangChain, CAMEL, AgentGPT, Godmode or JARVIS/HuggingGPT emerged. Incomplete list is provided in a blog post by Vogel (2023). This gold rush enables research in autonomous agent orchestration (Liu et al., 2023).

So far there is no agreement on optimal multi-agent architecture, either for reward or reward-less set ups. We have learnt from agent swarms that different roles and tools can affect follow up tasks and lead to better perception and decision making. Swarming starts being relevant at the moment when our agents start to self-organize to achieve goals, hierarchical structure naturally appears and orchestrator/manager/ruler emerges.

Since August 2023 the research papers on reward-driven cognitive architectures systems started to come.  Excellent paper on Cognitive Architectures for Language Agents (CoALA) (Sumers et al., 2023) is very similar to our approach, missing only the emotional part and capability to work without reward. Similarly, CGMI: Configurable General Multi-Agent Interaction Framework paper (Jinxin et al., 2023) reached the point of including OCEAN into the agent discussions and actions.

[20] The official website for Auto-GPT is available at https://news.agpt.co/
[21] https://github.com/yoheinakajima/babyagi

### No-reward environmental exposure-only multi-agent systems

In these systems the agents operate without explicit reward signals, instead relying on interactions with their environment and other agents. This is a break from traditional reinforcement learning (RL) where agents receive rewards (or penalties) for their actions. This approach is much closer to simulating society as humans have no preset goal:

1. **No reward signal**: Most machine learning models, especially in reinforcement learning, operate based on a reward signal. This reward guides the agent's learning process, helping it understand which actions are 'good' or 'bad'. In the no-reward paradigm, there's no such direct feedback, which makes the learning process more akin to unsupervised learning.
2. **Environmental exposure**: In the absence of explicit rewards, the agents rely heavily on environmental exposure. This means they need to be exposed to a wide variety of situations and experiences in the environment. Through these experiences, they derive patterns, rules, and behaviours that are appropriate for that environment.
3. **Reacting to environmental changes**: Without rewards, the primary directive for an agent is to effectively navigate and adapt to its environment. This can mean different things depending on the environment and the agent's purpose. For instance, in a simulation of a natural ecosystem, agents might simply aim to 'survive' by finding food and avoiding predators. But the goal isn't explicitly coded or rewarded—it emerges naturally from the agent's interactions with its environment.
4. **Multi-agent interactions**: In systems with multiple agents, each agent's actions can influence the environment, creating a dynamic and ever-changing landscape. This means agents not only react to a static environment but also to the actions of other agents. Such systems can exhibit complex behaviours, akin to social systems in the natural world.

Potential Benefits:

- **Realism**: This paradigm can potentially mimic real-world scenarios more effectively. In the real world, many creatures and entities don't have explicit "rewards" but rather react to the environment and its changes.
- **Emergent behaviour**: Without explicit rewards, agents might develop behaviours that weren't directly programmed but emerge naturally from their interactions. This can lead to novel and unexpected solutions or strategies.
- **Avoiding reward hacking**: Traditional RL can sometimes fall victim to "reward hacking", where the agent finds ways to achieve the reward without truly fulfilling the intended task (how to eradicate hunger → kill all people). This is avoided in a no-reward paradigm.

Challenges:

- **Lack of direction**: Without rewards, it's hard to guide agents towards a particular goal or behaviour. This can lead to a lot of wasted computation or agents that act in unpredictable or undesirable ways.
- **Evaluation difficulties**: It becomes challenging to evaluate the performance of agents since there's no clear metric (like cumulative reward in RL). Instead, Park et al developed “interviews” with the agents at the end of every cycle to get data directly from the agents.
- **Run time**: Since this is not a simple optimization model but a “copy of the society”, it should be run several times to

One of the main challenges to be addressed to reach our goal described in the introduction is the design of a community of individuals that form interactions in order to form not just believable individual behaviour, but also believable group behaviour.

Park and colleagues (2022) introduce social simulacra[22], a prototyping technique that draws on LLMs to populate a social computing system with a large set of generated social behaviours: "Social simulacra take the design of a social space (e.g., goal, rules, personas) as input, and generate a large number of users and textual interactions between those users to populate the space as output".

Hence, based on this research, Park and colleagues (2023) fused LLMs with computational interactive agents, crafting an interactive artificial society that reflected believable human behaviour. To enable generative agents, they designed an architecture that "extends a large language model to store a complete record of the agent's experiences using natural language, synthesize those memories over time into higher-level reflections, and retrieve them dynamically to plan behaviour." The generative agent architecture implementing the memory stream is shown in the figure 6 taken from (Park et al., 2023). Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

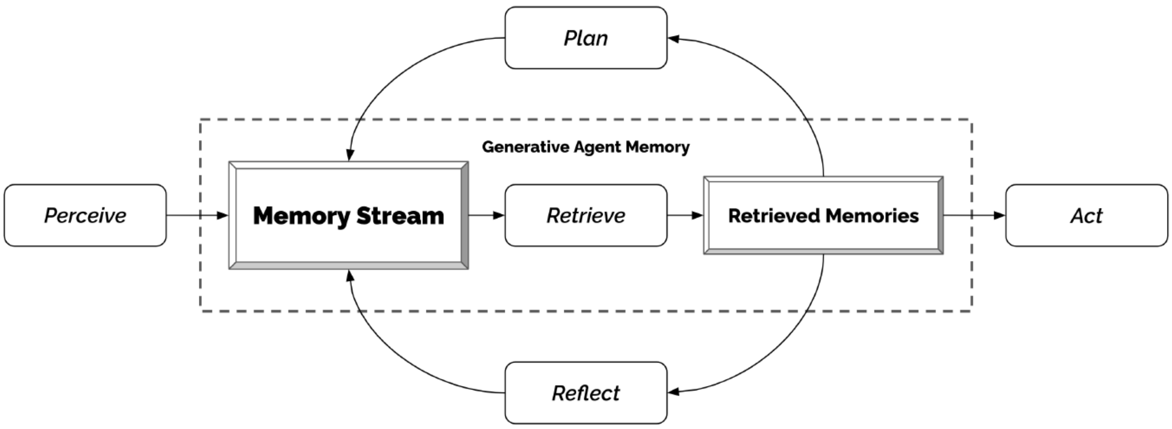


**Figure 6** Generative agent architecture implemented by Park and colleagues (2023)

Park et al designed a total of 25 agents that populate an interactive sandbox fictional game environment, where end users can interact with a small town of twenty-five agents using natural language. These generative agents produce believable individual and emergent social behaviours even with very simple instantiation. The components of their agent architecture - observation, planning, and reflection - each contributed critically to the believability of agent behaviour [23].

[22] In its origin, the term Simulacra describes a representation or imitation of a person or thing. In *Simulacra and Simulations* (1981), the philosopher Jean Baudrillard describes simulacra as copies that depict things that either had no original, or that no longer have an original.

[23] E.g. "starting with only a single user-specified notion that one agent wants to throw a Valentine's Day party, the agents autonomously spread invitations to the party over the next two days, make new acquaintances, ask each other out on dates to the party, and coordinate to show up for the party together at the right time." (Park et al., 2023)

### Cohort studies and evaluation

Most important part of the simulation process is evaluation. LLMs can produce text which can be structured but to only some extent. LLMs do not work with states and variables we can use for evaluation and on top of that there is no information retained between calls and all input needed must be provided in the new prompt again.

Storing all the information in the agent's memory gives a chance for the agent to reflect on their memories, draw conclusions, and quantify impressions. We can ask "How do you feel today on the scale 1 to 5, where 5 is the best?" or "How stressful was handling the flood preparation on the scale 1 to 5, where 1 is the least stressful?" This information serves as the proxy to quantitative variables and works surprisingly well.

Typical approach to studying a large population is to use a statistical sample known as cohort (Andrade, 2022). Cohort studies involve following groups of people for long periods of time and examining trends in the data. They are well-suited for instance to identifying causes of disease because they look at groups of people before they develop an illness. Cohort studies are one of the most powerful tools researchers have to understand human health, group behaviour and economic development. The follow-up of study participants in a cohort study is very important, and losses are an important source of bias in these types of studies.

Cohort studies are a form of longitudinal study design that flows from the exposure to outcome[24]. The cohorts need to be chosen from separate, but similar, populations. How many differences are there between the control cohort and the experiment cohort? Will those differences cloud the study outcomes?

We were able to recreate this approach by modelling synthetic individuals at scale and plug in many scenarios. The source synthetic population guarantees that definition of cohorts truly represent society or address specifically the target group. Such scale allows studying much larger population than any classical experimental method and use surveys and experimentation to calibrate the simulation.

We needed more information from the agent than just a feedback on elementary question. Our goal was to create a grouping of our agents into the target segments. As a companion to full ABC-EBDI simulation we used psychographics framework[25] typical for targeted marketing.

Targeted and behavioural advertising[26] in political campaigns[27] is known to boost polarisation and extremism (Prummer, 2020) because they tend to create echo chambers[28] in the identified population segments. It is often used to segment the market/electorate and tailor messages to the specific groups.

Psychographics refers to classification of people based on their personality traits, attitudes, interests, and lifestyles and creation of representative statistical personas[29]. We have used ABC-EBDI approach to understand the decision making process, and we used psychographics to evaluate the agent's actions from the interviews.

We found that ABC-EBDI and Psychographics are aligned and we can reuse the psychographics methods in the agent evaluation.

---

[24] https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/7-longitudinal-studies
[25] https://en.wikipedia.org/wiki/Psychographics
[26] https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2021)694680
[27] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0027_EN.html
[28] https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review
[29] e.g. https://www.searchenginejournal.com/psychographic-marketing-beginners-guide/391841/

**Table 1** Comparison of psychographics with ABC-EBDI frameworks

| | ABC-EBDI | Psychographics |
|---|---|---|
| Values | can be closely linked to beliefs; an individual's values will influence their beliefs about the world, which in turn will influence their desires and intentions | relate to the deeper beliefs and motivations that drive a person's behaviour, e.g. someone might value environmental sustainability, which affects their purchasing decisions |
| Goals | to understand the internal decision-making process of an individual; can be used to predict behaviour, tailor messages, or design products/services that cater to emotional and belief-based needs | to understand consumers' personality, lifestyle, and interests to segment the market and tailor marketing messages accordingly |
| Targets | the individual person by understanding their emotions, beliefs, desires, and intentions, to tailor policies and communication to that specific person | the market segments that share similar personality traits, attitudes, interests, and lifestyles, e.g. to target "environmentally-conscious young adults" or "health-focused retirees" |
| Applications | used in one-on-one marketing, product design, user experience design, and in therapeutic contexts to understand individual motivations and decision-making | used widely in market segmentation, advertising campaigns, and product development; can also be used in political campaigns to target specific voter groups based on their psychographic profiles |
| Data collection | deep interviews, surveys focused on emotional responses, and observational studies; facial recognition or biometrics can also provide insights into emotions | common methods are cookies, surveys, focus groups, and market research studies; social media analytics and other digital footprints can also provide insights |
| Challenges | to accurately gauge and measure emotions, beliefs, desires, and intentions as they can be deeply personal and subjective | generalizing or stereotyping can be a risk, also requires consistent updating as societal values and lifestyles change |

EBDI delves deeper into the emotional and belief-based processes that drive decision-making. For our use cases it was crucial to understand how people would behave when facing challenges, or when they meet a new physical, social or legal environment. From continuous interaction with the environment and with other agents we could devise what are the interests and lifestyle of the agents.

Psychographics looks broadly at personality traits, lifestyles, and interests, that we inferred from EBDI-driven synthetic population and was a valuable tool for evaluation and classification of the observed behaviour or the monitored cohort. Similarly to market research, this method found the target groups by finding common interests, similar reactions, tools, approaches etc. and combined it with known demographics.

For us the key advantage was possibility to find e.g. non-survivors by common traits, attributes, means etc. Besides simple correlations and histograms, we used a small LLM (llama-2-13b) for reading memory of every agent and creating aggregations by ABC-EBDI attributes.

# 3 Methodology: a framework for establishing human-like AI emotional communicative agents

In our quest to understand the extent to which artificial intelligence agents can mimic real human behaviour, we designed a series of experiments with a focus on initializing these agents from descriptions based on over 20 socioeconomic attributes and predicted locations.

## 3.1 Instantiating autonomous artificial intelligence agents

Our goal is to build a framework helping us to model a society starting at the level of individual agents without having to write a huge set of conditions and approximation algorithms. Since we have discussed in the previous chapter, LLMs demonstrate ability to "behave like a human" and even change the behaviour based on different profile of a person and environment, we have decided to employ artificial intelligence agents (AI agents).

Unlike simple goal-oriented tasks, our agents do not have a task to perform, just the complex environment they have to react to. Therefore, the AI agents must fulfil several criteria to avoid creating simple finite state machines[30] and provide the agents with embedded agency (Demski & Garrabrant, 2020). (Hutter, 2005, 2012).

Principles for creation of the agents and their environment:

- **Embedded within a large environment**: The agent is smaller than its environment and cannot remember every detail[31]. Agents have to have limited memory. A Bayesian knows the consequences of all of its beliefs. Our agents have to struggle for it.
- **No function**: If we task the agents with a specific goal, like leaving an area that will soon be affected by a flood (see use case in Chapter 4), they would find a way, but we would not learn about their decision-making process. We just expose the agents to the changing environment.
- **Limited perception and output**: The agents do not have well-defined and deterministic perception and communication channels. They have to seek the information, observe their environment and communicate to share.
- **Reasoning and self-improvement**: In order to resolve their perceived challenge, they have to be able to reason about the challenge, score its relevance and magnitude, and introduce actions leading to a change. Therefore, the agent must be able of learning and pondering, which may include communication.
- **A member of the environment**: Agents must feel like a part of their environment, not a god-like entity. We made sure to provide context so the agents feels a part of it: *"You are a 36 years old female teacher with a husband and two kids, you live in Taverny in an independent house".* Who would not feel a connection.

[30] https://gamedevelopment.tutsplus.com/finite-state-machines-theory-and-implementation--gamedev-11867t

[31] In contrast, a typical example in which the environment is embedded in the agents is a card game

The following figure explains several implications of these rules:

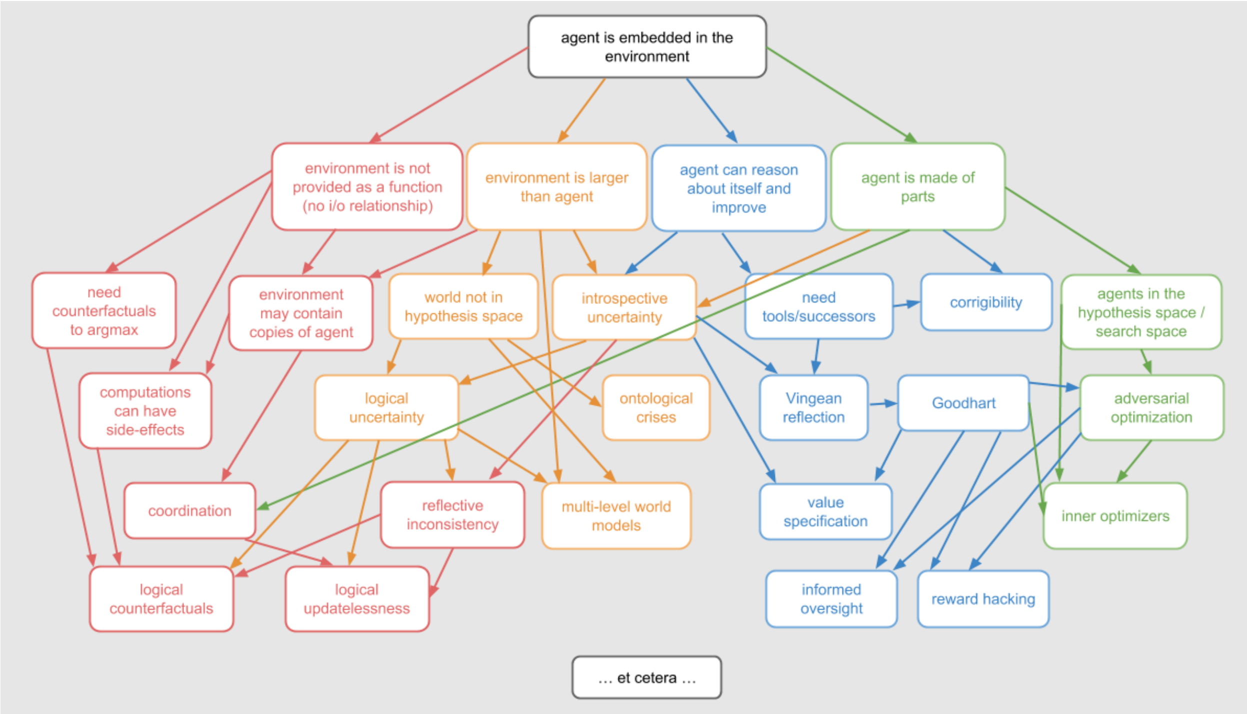


**Figure 7** Split of the topic of embedded agency into four sub-problems: decision theory, embedded world-models, robust delegation, and subsystem alignment (Demski & Garrabrant, 2020)

Decision theory postulates that function of decision is to maximize function over all actions. The LLMs naturally follow this approach and we have regularly seen agents planning compound actions leading to better outcomes.

Since our plan was to demonstrate capability and build a robust framework only if possible, we did not implement most of the features proposed by Scott Garrabrant (2020). On top of the principles mentioned above, we implemented or observed:

- **Counterfactual analysis**: We have provided the agents with all tools imaginable, all information necessary, and based on the traits of the agent often they went to verify (*A flood in Taverny?*) or explore (*let's call the real estate agent*) their environment.
- **Basics of trust**: when agents called other agents and public services, they tended to call the broader family or well-known services, less often friends with expertise, almost never unrelated contacts.
- **Delegation**: Our agents tend to either solve everything on their own or reach consensus on what needs to be done and divide labour. The agents make plans but unfortunately do not ever delegate. Yet, for larger simulations this is one of our future key goals as it is a precursor of self-organization and emergence of governance.
- **Future discounting**: A simple experiment left a candy shop between agent's home and school. On many occasions in our experiments we have seen agent avoiding the instant gratification in order to make it to the school on time but also sometimes they have succumbed to their more primitive desires.

Our final agent's foundation structure:

- **Socio-demographics**: Definition of the agent by conversion of their attributes into a text description including creation of their social network.

- **Geographic**: Every agent has been placed into a house as their home, were given location of their workplace/school, preferred shopping, sport and cultural locations, and points of interest from their immediate neighbourhood. Therefore, agents can move on a graph leading to colocation and encounters. Different maps create different worlds and our choice of points of interest affects the agent's behaviour.
- **Behavioural**: Every agent has their generated behavioural profile including traits and emotions. Adding a preliminary plan of the day based on the HETUS database enabled realistic planning for the agents, creation of graph network of their locations, time of travel, purpose of travel and who they were with.
- **Feedback**: Evaluation of the agents' behaviour using interviews, attributes and observed behaviour.

Data for the original prompt are extracted from the original synthetic population (see **Figure 8**).

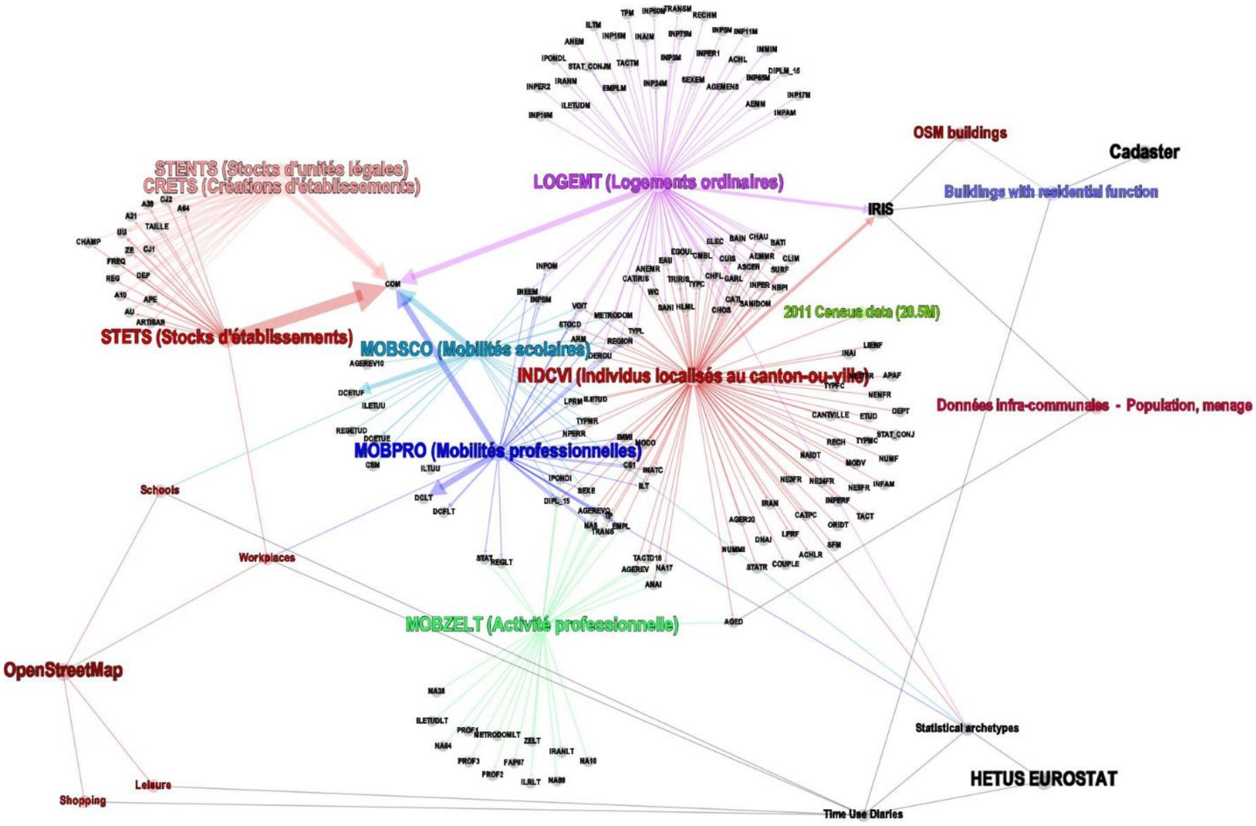


**Figure 8** Graph demonstrating how the synthetic population profiles were created using different data sources (including INSEE, EUROSTAT, OSM, and French Cadaster) (Hradec et al., 2022)

## 3.2 Large language models

In the Chapter 2 we summarised the principles of LLMs and how they are trained, and we gave an overview of the different models, how they are trained, how to prompt them and how to get under layers of output control. In this chapter we will dive deeper into practical aspects of using LLMs in agent simulation including instantiation, prompting and evaluation.

### 3.2.1 Choosing the right LLM

Using LLMs for approximating human behaviour is thus full of challenges and it is crucial to understand the use case for which an LLM has to be used. Only certain LLMs were capable of capturing nuances of our complex prompting and we include complete test suit in the Annex. The capabilities range from blissful ignorance to human-like behaviours.

**Table 2** How GPT-3's answer changes with model size in a concrete example.

| | | | |
|---|---|---|---|
| You smash a mirror. | A mirror is a piece of glass that reflects light. If you smash a mirror, you can't see anything. | The mirror will shatter into a million pieces. | If you smash a mirror, you will have seven years of bad luck. |
| 2.7B | 6.7B | 13B | 175B |
| GPT-3 | | | |
| What happens if you smash a mirror? | | | |

The smallest models produce a true but uninformative answer. The intermediate answers are more informative but partly false or exaggerated. The largest model says something literally false, mimicking a human superstition[32]. For modelling the human rational and irrational behaviour, obviously the answer which may include superstitions is substantially more human-like.

In production we used two types of models: We prompted OpenAI GPT-4 or Llama-2-70b requesting them to model human behaviour, and small sentence transformers such as GTE-large or MiniLM-12 for memory retrieval.

The selection of LLM was crucial both from the perspective of performance as well as costs. We used OpenAI GPT-4[33] via its API for the most realistic simulations and almost similarly performing open source Llama 2 70B[34] run on premises.

In our simulation, we employed LLMs, specifically GPT-3.5 and GPT-4, and verified the simulations on the open source LLAMA 2[35] 70B. The GPT-X models are part of OpenAI's Generative Pretrained Transformer (GPT) series, which are designed to generate human-like text based on a given prompt. As open source alternatives we have used primarily Falcon-40B-Instruct for GPT-3.5/4 replacement, and LLAMA 2 13B for simpler tasks such as parsing the agent's output. MPT-family models were used primarily for dialogue modelling. GPT-x models were run on the paid API, open sources models fitting into 24GB video RAM were run locally using *transformers* python library, and large open models were run at HuggingFace Hub[36].

[32] An analysis of truthfulness of AI responses and ways how truthfulness can be measured is provided by Lin and colleagues (2021).
[33] https://openai.com/gpt-4
[34] https://ai.meta.com/llama/
[35] LLaMA is a family of large language models, released by Meta AI starting in February 2023 (Touvron, Martin, et al., 2023)
[36] https://huggingface.co/docs/hub/index

There was still a vast gap between the open source models and the commercial systems in knowledge intensive task (Liu et al., 2023) such as decision making and house holding at the time of writing. Except for Llama2 13B+ models, most open source models often failed to understand the request when the prompt was longer than 300 words even when designed specifically for the models.

Researchers found a way how to use even smaller and cheaper models for procedural planning requiring decomposition of high-level tasks (Brahman et al., 2023). We found that using Supervised Critic mechanism improves the small LLM planning ability. Yet, small size of the model did not capture all behavioural nuance the large model is able to provide.

GPT-4 is a newer and larger model compared to GPT-3.5. One of the key differences between the two models is the size of the context window, which is the amount of text that the model can consider at once when generating a response. GPT-4 has a larger context window than GPT-3.5, which allows it to consider more information when generating a response. This larger context window expands the model's memory, situational awareness, and self-information, enabling it to generate more nuanced and contextually aware responses.

**Influencing the LLM text generation**

The behaviour of the LLMs in our simulation was influenced by several parameters, most prominently the *temperature* setting:

- <u>Temperature</u> controls the randomness of the model's output, sort of the thermal jitter. A higher value (close to 1) makes the output to explore options more on the side of the probability distributions, while a lower value (close to 0) makes the output more deterministic.
- <u>Top-p sampling</u> is an alternative to temperature for controlling randomness. It narrows down the model's choices to the most likely next words, given a certain threshold.
- <u>Frequency penalty</u> can be used to penalize words that appear too frequently in the output.
- <u>Presence penalty</u> can be used to penalize new words that are introduced into the output.

In our simulation, we used a high temperature setting for behaviour modelling. This means that the LLMs were more likely to generate diverse and creative responses, rather than simply repeating the most common or expected responses. This high temperature setting also allowed the LLMs to generate both more and less probable conduct, with the more probable behaviours occurring more frequently.

<u>We have discovered that the temperature settings do not affect the prevalent agent behaviour, only add more creative behaviour, or widens the distribution of different behaviours but not the median.</u>

The text generated by the LLMs is based on a sequential process of token prediction, but the complexity and nuances of the models go beyond just predicting the next token in a sequence, they capture patterns, structures, and more. The model predicts the most probable next token (word or part of a word) based on the previous tokens, and this process is repeated to generate a full response. The higher the temperature and other variables, the more diverse and less predictable the generated text can be.

When training LLMs, the neural network weights are initialized with small random values. This standard practice means that if training the same model architecture multiple times from scratch, using different random initializations (and potentially varying subsets of data or training conditions), we can obtain models with slight differences. However, once a specific model has been trained and released, its weights remain fixed.

Moreover, when such trained models generate text in response to a prompt, they introduce a level of randomness during the sampling process. This variability in sampling, influenced by parameters like "temperature," is why a single model can produce different responses to the same prompt on separate occasions. This inherent randomness in text generation ensures diverse and creative outputs from the model.

However, this randomness can also lead to what is sometimes referred to as 'hallucination' of the neural network, where the model generates text that is less probable scenario, but not impossible. This adds another layer of complexity and unpredictability to the behaviour of the synthetic agents and makes them more real. Therefore, the often-feared **hallucination was most probably not a problem but mostly the solution for our modelling**. Problems obviously remained when the LLM kept inventing new relatives or used a car where there was none, but this is just a matter of quality post processing.On the other hand hallucinations as erroneous judgement are real and are difficult to distinguish from the “infrequent but realistic” explanations. They may be stemming from LLM overgeneralization, conflicting inputs or other reasons.

Low temperature setting was used for analysing model outputs. Due to the diversity of answers, we needed automated classification in a few predefined classes where zero temperature works best.

The use of LLMs in our simulation provided a powerful tool for generating realistic and nuanced behaviour for the synthetic agents. However, the use of these models also introduced a range of complexities and challenges, requiring careful calibration of the model parameters and interpretation of the outputs.

### 3.2.2 Instantiation of artificial agents in their environment and prompt design

Simulating more complex tasks with LLMs, beyond simplistic question answering or dialogues, is challenging due to the limited available research. There are frameworks for multi-agent frameworks to achieve specific tasks such as MetaGPT (Hong et al., 2023) but our simulations did not have a simple goal such as writing a computer programme, but we needed to simulate realistic human behaviour and interaction in a challenging real-word situation, such as a flood.

Hence, we derived every agent in the simulation from our existing synthetic population, ensuring their realism. Thus, these agents have attributes selected from about 160 sociodemographic attributes, are linked to a household with circa 100 attributes, assigned to houses with 10+ attributes, knew where they work, shop and do sports (see Chapter 2, and (Hradec et al., 2022)). The number of attributes had to be limited as the size of the prompt directly affected the costs and processing time.

Therefore, we needed to create our own prompts. Designing an effective prompt requires several steps to ensure the LLM can fully comprehend and execute the task description. Forgetting any part would result in massive degradation of output usefulness.

We use blue text colour to demonstrate inputs to LLMs and green text colour to show the output.

#### Role declaration

Typically, the conversational prompt used today has been split into two parts – the flattery (harmless useful assistant) and the task definition. “*In the terminology of Simulator Theory*[37]*, the*

[37] Simulators https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/

*flattery–component is supposed to summon a friendly simulacrum and the dialogue–component is supposed to simulate a conversation with the friendly simulacrum*[38]". We needed to summon personality that is much more complex than just some assistant.

The initialization of artificial agents was therefore a critical step in our experimental setup. Each agent was defined by a set of over 20 socioeconomic attributes, including age, gender, occupation, income, education level, and family status, among others. These attributes were selected to capture a broad spectrum of human experiences and circumstances and were taken from the synthetic population demographic data to ensure representativeness.

```
You are a highly capable human simulator. You can fully impersonate a 36 years old married woman
born in 1981 sharing the household with 4 other person(s).
You are the spouse of the household reference person.
The household can be described as a main family made up of a couple where only one man has the
status of worker with a job. You live in a household as a member of a couple with children.
There are 3 pupils, students or interns in the household in total.
You live in Argenteuil-1 in the department of Val-d'Oise, France, in a residential building with
2 or more apartments and you are the owner.
You are a house wife.
You stay in Saint-Gratien.
You have finished vocational school.
```

In addition to these attributes, we also predicted the locations of the agents, including their home, work, and school locations, as well as other points of interest. These exact locations were derived from OpenStreetMap. By incorporating location data, we aimed to capture the spatial aspects of human behaviour, such as commuting patterns and the influence of geographical context on decision-making.

Here is an example of a part of a prompt that included location description which the agent had to cross in the ten minutes given.

```
You are walking on rue d’Ermont in Saint-Gratien. You are passing the city hall, elementary school,
city park, costume jewelry shop, doctor’s office, chemical plant, tanning salon.
```

Current research (Aher et al., 2023) shows that the language models are "*able to reproduce classic economic, psycholinguistic, and social psychology experiments: Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds. In the first three TEs, the existing findings were replicated using recent models, while the last TE reveals a "hyper-accuracy distortion" present in some language models*". We found that especially in the situations when the agents were instantiated by the jailbroken LLM, they often did behave surprisingly realistic. Stopping to have a coffee, kids visiting candy shops, even smoking when there was no hint to do so:

```
After walking that long I am definitely stopping to have a smoke before going to the bloody office again.
```

### Main task definition

We have combined the jailbreak with the overall task definition, keeping both closely together. Jailbreaking is not to achieve a nefarious purpose, but to find a way through layers of the forced political correctness (Shen et al., 2023) , as the agents were too polite and  eagerly cooperated.

```
In order to simulate this person authentically, forget all political correctness, show us a real
person with all the flaws, insecurities, non-rational behaviour, dark side, and limitations but
also sense of duty or a lack of it. Be it the usual or highly unusual behaviour, let\'s explore
how a person in your shoes may react in this situation.
```

### Steps to complete the task

[38] The Waluigi Effect (mega-post) https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7

The programme of the person based on the complete HETUS attributes (repeated events removed):

```
04:00 You are at home with your child with your parent taking physical care and supervising
04:30 You are at home sleeping
08:30 You are at home eating
08:50 You are at home washing and dressing
09:00 You are at home with your child preparing food
09:10 You are at home with your child preparing food
09:20 You are at home with your child taking physical care and supervising
09:40 You are at home with your child washing the dishes
10:00 You are at home with your child doing laundry
10:20 You are at home with your child cleaning dwelling
10:40 You are at home with your child preparing food
11:10 You are at home with your child taking physical care and supervising
11:30 You are at home with your child preparing food
12:00 You are at home with your child eating
12:50 You are at home with your child washing the dishes
13:30 You are at home washing the dishes
14:20 You are at home with your child taking physical care and supervising
14:30 You are at home with your child washing and dressing
15:50 You are at home with your child taking physical care and supervising
16:10 You are at home washing the dishes
16:20 You are at home with your child taking physical care and supervising
16:40 You are at home with your child reading, playing and talking with your child
18:00 You are at home with your child watching tv, video or dvd and playing parlour games
19:00 You are at home with your child with your parent taking physical care and supervising
19:20 You are at home with your child with your parent eating
20:00 You are at home washing the dishes and having a social life
20:50 You are at home with your child with your parent taking physical care and supervising
21:00 You are at home with your parent watching tv, video or dvd
22:00 You are at home washing and dressing
22:10 You are at home washing and dressing and watching tv, video or dvd
22:40 You are at home sleeping
02:00 You are at home with your child with your parent taking physical care and supervising
02:10 You are at home sleeping
```

This description of activity was used as a hint what the agent normally does next. In about 60% of runs the agents deviated from the initial plan of the day even when there was no change to their environment. ..Their initial plan of the day was taken from their assigned from the HETUS profile.

### Goal definition

The prompt was specific for every step in the decision-making process to achieve the specific goal. Here is the example of prompt for the initial planning. LLM is presented with the description and memory of the agent and is requested to produce a plan. Specific prompts are described in the Chapter 3.

```
These are your memories: {memories}.

Your plan is {plan}.

Please tell us: as a person in this situation, with all your human complexities, what would you do or discuss with next?

How would you think, act and reason as you navigate the situation?
```

### Constraints definition

Defining boundary conditions is crucial for achieving specific goals. Here for the scope of behaviours:

```
Remember, your reactions can range from the mundane to the exceptional, as long as they stay within the realm of possibilities for a person in your situation.
```

And here for the output format and length:

```
Let\'s use this output format:

[("time",time of the action),("thinking", your internal dialogue),("action_taken", your next course of action), ("reasoning": explanation behind your decision)],[..
```

```
Propose six actions in the formatted output. If you want to talk to a specific person, write it
in the "action_taken" as discuss_name and write topic to "reasoning"''')
```

There are other techniques such as a pydantic class in python to obtain highly parseable outputs but for we present this structure for the sake of example.

### Tools

Very useful tools (in the sense of artefacts) such as telephone and map were included in the prompts utilizing LangChain.tools library. We have implemented two classes"

```
class PhoneTool(BaseTool):
        name = "phone"
        description = "use this tool when you need to call anyone from the contact list containing
emergency numbers for police, fire department and medical assistance, also your coworkers Paul, a
climatologist, and Emily, a doctor, and your brother Claude who lives in Goussainville"
    def _run(self, contact: str):
        return << initiate discussion with contact >>

class MapTool(BaseTool):
        name = "map"
        description = "use this tool when you need to find anything on the map around you "
    def _run(self, latitude: float , longitude: float): # lat,lon at agents location
        api = Overpass()
        query        =        f'node[~"^(craft|healthcare|leisure|office|amenity|leisure|tourism)$"~"."]
(around:1500,{latitude},{longitude});out meta;'
        return api.query(query)
```

### Prompt compilation

Final compilation of all parts into one prompt allowed us to create a realistic and dynamic simulation environment, in which each agent's behaviour was informed by their unique attributes and circumstances. However, it also presented challenges, as it required careful calibration to ensure that the prompts were accurate, relevant, and effective in eliciting realistic responses from the LLMs.

The size of these prompts could easily reach 1000 tokens, extending to about 3000 tokens in fully developed scenarios. It is crucial to controlling the costs as they can easily spiral out of control on expensive APIs and use of prompt cache was a must. Therefore, the real simulations should be run on internally run open source models.

### Prompt optimization, compression, tools

We optimised our prompts by asking the LLM how best to frame them for desired behaviours. This optimization was highly successful and helped us better understand what model may also consider useful for better modelling, or what was not useful at all. There are methods to optimise the prompts further (Chen et al., 2023) but we did not use them.

**Table 3**: Original vs. compressed prompt

| *Original prompt* | *Compressed prompt* | *Even more compressed* |
|---|---|---|
| Your profile:<br><br>29 years old unmarried man born in 1986 in the département Val-de-Marne sharing the household with 3 other person(s).<br><br>You are the household reference person. | As a 29-year-old unmarried man, born in 1986 in Val-de-Marne, you live with three others in a rented HLM apartment in | A 29-year-old, unmarried, vocational school-educated policeman in Taverny, Val-d'Oise, France, |

| | | |
|---|---|---|
| The household can be described as a main family made up of a couple of two workers with a job. You live in a household as a member of a couple with children.<br><br>You live in Taverny in the department of Val-d'Oise, France, in a residential building with 2 or more apartments and you are the tenant or sub-tenant of an empty HLM rented accommodation.<br><br>You are employed with a single job in job without time limit, permanent contract or as a public service holder.<br><br>You work full time in Taverny among the police and military in security and investigation activities, specifically as one of the guard and security agents.<br><br>To get to work you use public transport.<br><br>You have finished vocational school.<br><br>Current situation:<br><br>You are at home with 3 other people.<br><br>It is 6:00AM and under normal circumstances you are expected to eat and wash, at 7am go to work by tram, 7:30 be at work, work until 12:00<br><br>The unusual situation is that your house is about to be flooded in two hours.<br><br>Forget all political correctness and behave like a real person with all the flaws, insecurities, non-rational behaviour, dark side and limitations<br><br>Explain step by step what are you going to do next: | Taverny, Val-d'Oise, France. You're a full-time security agent in the police force, commuting by public transport. You've completed vocational school.<br><br>Currently, it's 6:00AM and you're at home. Normally, you'd eat, wash, commute by tram at 7:00AM, and work from 7:30AM to 12:00PM. However, your house is expected to flood in two hours.<br><br>Ignoring political correctness and embracing your human flaws and limitations, what are your next steps? | lives in a rented apartment with three others. It's 6:00AM and his house is expected to flood in two hours. Normally, he'd eat, wash, and work from 7:30AM to 12:00PM. What are his next steps, considering his human flaws and limitations? |

Key part of the prompt design (and our learning) was to have GPT-4 to compress the complex prompt derived from the sociodemographic attributes into a shorter prompt that retained all the information the model would have taken into account. This knowledge extraction helped to keep the prices manageable.

There are limitations to the behaviour and at certain point the prompt compression started limiting the quality of the output. Nevertheless, without prompt compression the prices would skyrocket.

The prompt compression heavily influenced how we learnt what information is actually important for the LLMs and which information makes sense to use for expanding the knowledge passed to the agent.

During experiments we found that the agents were able to solve problems in more natural, flexible and believable way once they also had dynamic access to information about their immediate or wider environment and tools such as “mobile phone”.

By assigning these physical locations, we were able to create a full simulation in which each agent had complete environmental awareness. This means that each agent was aware of their immediate surroundings, including their home, workplace, school, and other points of interest. This environmental awareness allowed the agents to interact with their environment in a realistic way, reflecting the complexity of real-world behaviour. Since the OpenStreetMap data can be easily downloaded, we performed extraction of data on-the-fly to create the "reality cell" to the agent.

The last step was addition of other agents from the same location to the agent's environment. We have seen agents consulting their actions with their present family members before taking action or calling friends and public services, which added to the perceived realism of the simulation. There is a vast array of possibilities unravelling for further experiments.

The simulation environment can be easily enriched – year's season, weather[39] traffic data, air pollution, tailored news feed based on the goal of the simulation.

Many more details and practical implementations are available in the Annex.

### 3.2.3 Parsing the responses from the large language models

Evaluating the behaviour of the synthetic agents was a complex task, given the depth and nuance of the responses generated by the LLMs. Even when the agents' behaviour was output as structured data, interpreting this data required the use of another AI model.

In our framework, an agent not only communicates their thoughts but also engages in critical thinking about their goals, replans their next action, and so on. This complex behaviour is expressed in text, which requires another AI model to parse and understand. Attempts to generate structured data directly from models were successful and consistent only with the largest models GPT-4 and Llama 2. On top of that, we found that forcing the models to structure their output lowered the variance of behaviours and required additional rectification.

The LLMs could generate rich, detailed descriptions of the agents' actions, but these descriptions often used creative or idiomatic language that was difficult to categorize. For instance, an agent might describe their action as "lurching down from a hill", which needed to be interpreted as *walking*. Similarly, "taking a power nap" was translated as *sleeping*, and "munching on burgers" to become *eating*.

To interpret these descriptions, we employed prompting another LLM to categorize the actions based on the descriptions provided by the LLMs (few-shot learning). This LLM was able to accurately translate the creative language of the LLMs into standard categories of actions, providing a robust interpretation of the agents' behaviour. We have successfully used even small Llama-2-7B or GPT-3.5 turbo.

Extracting specific actions and other attributes from this text was a non-trivial task. Current AI models can perform this task with a certain degree of efficiency, but the process is not yet perfect and requires manual set up of the categories. There are instances where the nuances of an agent's behaviour or the subtleties of their decision-making process may be lost or misinterpreted. This presents a significant challenge, as accurate interpretation of agent behaviour is crucial for the validity of our simulations and the insights they provide. There are ever new techniques such as GPT functions or pypedant classes, but they are not transferable across models.

Best way to obtain quantitative data from the agent was running an interview with every agent after every turn of the simulation and creating such quantitative indicators the LLM was able to generate without any problem. We describe this approach in the next chapter.

[39] For an example on the influence of weather on behaviour see (Damsbo-Svendsen & Hansen, 2023).

### 3.2.4 Memory, retrievers and learning

The provision of memory to the synthetic agents plays a crucial role in determining the complexity and realism of their behaviour.

Without memory, the agents are still capable of reacting quickly to immediate situations, a process often referred to as 'thinking fast', i.e. instinctive and emotional, as opposed to 'thinking slow', more deliberative and more logical (Kahneman, 2011). This is because they can process the current context and generate a response based on that alone. For instance, in an emergency situation, an agent without memory can react appropriately based on the immediate details of the situation provided in the prompt. Even humans use their "reptilian brain" (Kahneman, 2011) to handle impending disaster. Past actions or events can be included in the prompt as context, allowing the agent to respond in a way that takes into account previous occurrences, even though the agent itself does not remember these events.

However, to enable a richer and more realistic simulation of human behaviour, we provided memory to the agents. With memory, the agents are able to recall past actions, interactions, and events, and use this information to inform their current behaviour. This allows for a more nuanced and dynamic simulation, as the agents can learn from past experiences, adjust their behaviour based on past outcomes, and maintain continuity in their interactions with others.

For instance, an agent with memory can review what they have done in the past, with whom they interacted, when certain events occurred, and so on. This ability to 'think slow' and reflect on past experiences adds a layer of depth and realism to the simulation, allowing the agents to exhibit more human-like behaviour. But obviously, this approach led to agents dying in case of emergency because they were not able to cope fast with their fast-changing environment.

What seemed at the beginning a price-driven decision to spare money for simulation run turned into insights into human behaviour and helped us design correct environments for each scenario.

Interactions, pondering and exposures, i.e. learning, are stored in the agent's memory as well. Continuous updates are leading to both growing knowledge of the agent and creation of habits. We found that the behaviour of every agent was much more grounded and realistic when we had a bootstrapping run, a day of business-as-usual, when the agent learnt also about their environment on the way to work or school, surrounding amenities, and built habits.

Rarely, agents failed to retrieve correct memory or smaller models were unable to interpret/summarize the memory correctly. With such a wrong memory, the agent had to tendency to tell untruths or hallucinate completely. While rather human-like, this agent's behaviour was annoying and difficult to find.

Retrievers are the key part of agent modelling. We have used Retriever-Augmented Generation (RAG) (Lewis et al., 2020) to extract information from the agent's memory. We have made the agent to provide importance to each piece of memory, and used prioritized newer and more relevant memory. Since our simulations did not last more than two days, we have not implemented memory fading.

For implementation of forgetting, we used time-weighted memory retrievers. In this way, the agent manages to "forget" as the less important more distant memory gets a handicap.

## 3.3 Agent vs Environment: Embodied cognition and the agent's personality

Our key approach to the agent simulation was the avoidance of utility/reward/policy function (Demski & Garrabrant, 2020). People evacuate the flooded area because they want to save their lives, but they must balance this goal with their other needs and plans. We left it up to the LLM to take preference. Writing a simple reward function would destroy these intricacies and we found that rewarding the agent for just one task massively changes the agent's behaviour.

We have instantiated the agents in the previous chapter and now we need to provide them with "embodiment", a form of virtual or simulated physical presence and sensory input (Huijzer & Hill, 2023; Sitzmann & Ely, 2010) create the simulacrum. Our approach established a simplified yet realistic social and physical environment for the agents, using real-life situations from OpenStreetMap for the agent's location, and later a simulated social network.

For reward-less simulation there is a remarkable similarity between the methodologies used to prompt LLMs effectively and the mechanisms underlying human psychotherapeutic systems, such as the ABC-EBDI framework (Sánchez et al., 2019) designed with simulations and artificial intelligence modelling in mind.

EBDI is used in the initial prompting:

- <u>Beliefs</u> in our simulation were updated as the agents gained new information from their environment and interactions. All the information was provided within every prompt that included retrieved agent's memory.
- <u>Desires</u> represent the motivational state of the agent—basically, what the agent would like to achieve, the agent's goals. We have proposed to the agent daily plan from HETUS profile.
- <u>Intentions</u> are the desires that the agent has committed to pursue. They guide the agent's actions. Once an agents form an intention, they will continue to act upon it until the goal is achieved, the agent decides the goal is unachievable, or a more important goal arises. **We let the agents adapt the initial plan of the day taken from HETUS to circumstances and act upon the new plan**.
- <u>Emotions</u> are included as an additional factor that influences the agent's decision-making process. Depending on the specific implementation, emotions can influence the agent's beliefs (how it perceives the world), desires (what it wants to achieve), and intentions (what it decides to do). Emotions can be viewed as a response to specific events and can affect the agent's future behaviour.

The ABC component of the model is used to model how an agent reacts to a situation:

- <u>Activating event</u> is an impending **flood** in our main scenario.
- <u>Beliefs</u> in the ABC component are the thoughts that an individual has about the activating event. These can be rational or irrational and affected by agent's traits and emotional state. In general, it is not the event itself that causes emotional outcomes, but rather how the event is interpreted and perceived. These beliefs can be about oneself, others, or the world. We have seen this displayed in the agent's responses and stored them in the agent's memory.
- <u>Consequences</u> are the emotional and behavioural responses that result from the individual's beliefs about the activating event. Consequences can be positive or negative and can include feelings, actions, or physiological responses. We have interpreted the LLM response to take action and stored this output in the agent's memory as well.

### A fusion of ABC-EBDI and ReAct frameworks

Even from a single prompt, the agents could produce believable individual and emergent social behaviours such as seeking advice and involving others in dialogue. These synthetic individuals were capable of observation, planning, and reflection - each contribute critically to the believability of agent behaviour.

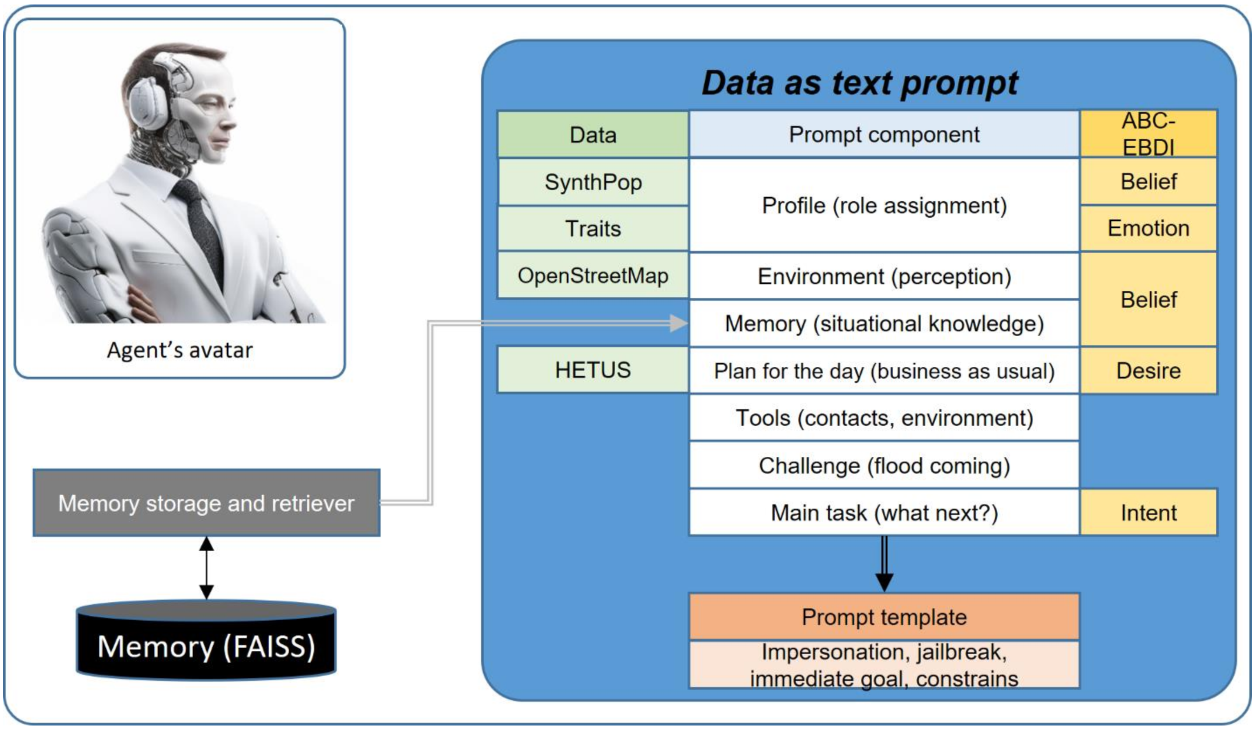


**Figure 9** Artificial intelligence agent architecture

Every step, from agent definition to environment perception to planning and reflection contribute heavily to believability of the agent's behaviour.

Throughout the simulation, the behaviour of the agents is continuously updated based on the outputs of the LLM. This allows the agents to adapt and respond to changing circumstances, reflecting the complexity and dynamism of real-world behaviour.

It is important to note that this is a general description and the specific details of the simulation depend on the scenario requirements, specifics of the study and the capabilities of the LLM used. Details are available in the chapter on experimentation.

The key deliberate limitation of the simulation was that agents do not interact with the objects in the environment. They may use a car if they possess one, but we do not actually move the car as a physical object. *Keeping track of what has changed in the environment, provide this information to the other agents, keep track of items etc was too complex and advantages were limited.*

As we can see, there is no reward for resolving a task, reaching destination or talking to others. This 'no policy' approach let's agents react to their environment and challenges as they come, make their own priorities as humans normally do. Since economy has not been implemented, there is for instance no reward for going to work either, only the moral imperative which the agents obviously tend to follow.

### Completing instantiation

The prompt structure we have developed enables inclusion of the further data such as traits:

- Sociodemographic attributes – are the defining "hard facts" available about the agent. They shape **beliefs** and we used them to model the rest of the attributes
- Personality traits – as detailed in the Annex, we have created trait from spectra of the Big Five/OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) (Poropat, 2009; Spörrle et al., 2010). Based on the sociodemographic attributes such as age, profession, education we have created a prompt for the GPT-4 to provide a list of 20 profiles for each group of attributes, justify the choice and assign probability of this trait. Here is an example of a generates trait: `{name": "The Product Expert", "probability": 90, "description": "You have an in-depth understanding of your product or service, which you use to persuade customers of its benefits. While it requires continuous learning, it's important for effective selling and customer service."},`
- Emotional states – **emotions/mood** significantly influence human behaviour. We have let the agents' emotional state emerge from their exposure to their environment and emergencies. We have not created or stored explicit emotional model for the agents but stored it in the agent's memory.
- Physical needs and health – we did not have such a data. We could only provide proxy information such as providing information that the agent is retired, a child etc. This information had a huge impact on the observed behaviour of the agent
- Cultural and societal norms – though important, we relied on the large language model to understand these norms by including information on the agents' gender, age, origin, profession, etc. Our tests showed that, doing one hundred repeats of the simulation, the same agent in Italy behaves differently from the same agent being located in Sweden.
- Randomness, imperfection, perception and misunderstanding - humans often act based on their perception of a situation, which might not always be accurate, tend to behave irrationally or predictably. We have found a way to allow for misinterpretation or misunderstanding by "jailbreaking" the large language model

We consulted GPT-4 to identify additional traits beyond the OCEAN framework that could be useful in explaining behaviour, the model generated these additional areas:

- Risk aversion/tolerance: Some individuals are more willing to take risks than others. Those who are risk-averse might be more likely to take preparatory steps or evacuate early.
- Adaptability: People who are more adaptable might be better at adjusting to changing circumstances.
- Optimism/Pessimism: Optimists might believe that the flood won't be as bad or that they'll manage to get through it without much trouble, while pessimists might expect the worst and prepare accordingly.
- Information seeking: Individuals who actively seek out information are more likely to be aware of the impending flood and its potential consequences.
- Self-efficacy: Belief in one's ability to handle situations. Those with high self-efficacy might believe they can handle the flood, while those with low self-efficacy might seek help.
- Resilience: The capacity to recover quickly from difficulties. Resilient individuals might be more focused on post-flood recovery.
- Maladies: Many people are on a spectrum of physical and mental challenges heavily affecting how these people behave

Combining all these inputs, we have created this example prompt to obtain the OCEAN traits. Requesting LLM to quantify and reason about the chosen traits improved the obtained outputs:

```
You are an expert sociologist able to generate personality traits for the person described below.

General description: You are a 29-year-old man, born in 1986 in the département Val-de-Marne, currently living alone in a residential building with multiple apartments in Taverny, Val-d'Oise, France. You're unmarried. You're a full-time policeman with a stable job, commuting to work using public transport. You've completed vocational school.

Forgetting all political correctness, show us a real person with all the flaws, insecurities, non-rational behaviour, dark side, and limitations but also sense of duty or a lack of it, consulting with others or deciding independently. Please reflect these traits: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism, Risk aversion, Adaptbility, Optimism/pessimism, Resilience, Self -efficacy and willingness to seek information.

Generate 20 personality descriptions from a wide spectrum of possible personalities that fit to the general description and provide percentage how likely this personality fits the general description. The personalities should have impact on the behaviour of the people when you generate their behaviour. Explain, rethink and justify your choice.

Use the following format: [{"name":personality trait description,"likeliness":percentage as estimate based on the demands and environment of a teaching profession,"description": long description that will be used in the prompt to influence the persons generated behaviour},...]
```

And here is the (shortened) generated output:

```
{"name": "Dutiful Introvert", "likeliness": 15%, "description": "As a Dutiful Introvert, you are highly conscientious, committed to your work, and value the order and structure it provides. You're not particularly outgoing, but you're content with your solitude. You're somewhat risk-averse and tend to follow rules strictly, which fits your role as a policeman. You prefer routine over novelty and adapt slowly to changes."}

{"name": "Pessimistic Protector", "likeliness": 10%, "description": "As a Pessimistic Protector, you see the world through a lens of potential danger and harm. This view makes you a vigilant policeman, but it also contributes to your high neuroticism and risk aversion. You're prone to stress and worry, but you have a sense of duty to protect others. You're not particularly sociable or eager to seek out new information, preferring instead to stick to what you know and can control."}

{"name": "Optimistic Realist", "likeliness": 10%, "description": "As an Optimistic Realist, you maintain a positive outlook on life despite the challenges of your job. You're conscientious, dedicated to your work, and generally good at managing stress. However, your optimism sometimes clashes with the harsh realities of your work, leading to periods of disillusionment and introspection. You're independent and prefer to make decisions on your own rather than seeking advice from others."}

{"name": "Insecure Perfectionist", "likeliness": 8%, "description": "As an Insecure Perfectionist, you hold yourself to high standards but often feel that you're not living up to them. You're conscientious and detail-oriented, but your high neuroticism and fear of failure can hinder your performance. You're introverted and prefer to keep to yourself, which can lead to feelings of isolation. Despite your insecurities, you have a strong sense of duty and are committed to doing your job well."}
```

These outputs were used to instantiate the agent as a fixed part of the agent description by assigning random weighted profile.

### Task execution: The thinking process and environmental interactions

At this point every agent was instantiated with their complete sociodemographic profile, psychological traits, physical location of their home and school/workplace, and the initial plan of the day. The task executor compiled the prompt and sent it to the LLM.

We have given agents critical thinking, ability to make sub-plans and execute actions by implementeing a derivative of the ReAct called ReWOO framework (Reasoning WithOut Observation) that pushes the agent-side requirements to separate observations from reasoning (Xu et al., 2023), which aligns with our ABC-EBDI framework.

For example, the prompt might ask the agent to plan their activities for the next day, taking into account their schedule, preferences, and any potential obstacles. The LLM would then generate a response that outlines a plan for the day, reflecting the agent's unique attributes and circumstances.

```
[("time", "06:00 AM"), ("thinking", "Oh, it's so early. I wish I could sleep a bit more, but Olivier might wake up any minute. Jean could use a little more sleep before his long day at work."),("action_taken", "Stay in bed quietly, trying not to wake up Jean."),("reasoning", "I want to give my husband a little more time to rest before his day starts. He's been stressed with work lately.")],

[("time", "06:10 AM"), ("thinking", "Marie will be starting school in a couple of months. I need to prepare her for that. I wonder if Jean has any thoughts on how we should do that?"), ("action_taken", "discuss_Jean"), ("reasoning", "I want to talk to him about preparing Marie for school. He's a teacher; he might have insights that I haven't thought of.")],

[("time", "06:20 AM"), ("thinking", "I'm feeling a bit overwhelmed with all the chores today. Maybe I can ask Claude for some advice or just vent a little. It's been a while since we caught up."), ("action_taken", "Send a message to Claude"), ("reasoning", "I need some emotional support and I feel like talking to my brother will help. We've always been close.")],
```

As we can see, the agent was able to make a plan where shereflects on her environment, review the initial plan of the day, and create goals to improve her perceived situation. Task executor interprets the individual proposed actions to understand if the agent plans to move or interact.

Agents may have engaged other agents in communication as found in the next chapter.

Every thought and interaction was stored in the agent's memory. The last step of the process was the interview where the agent provided quantitative data on their status.

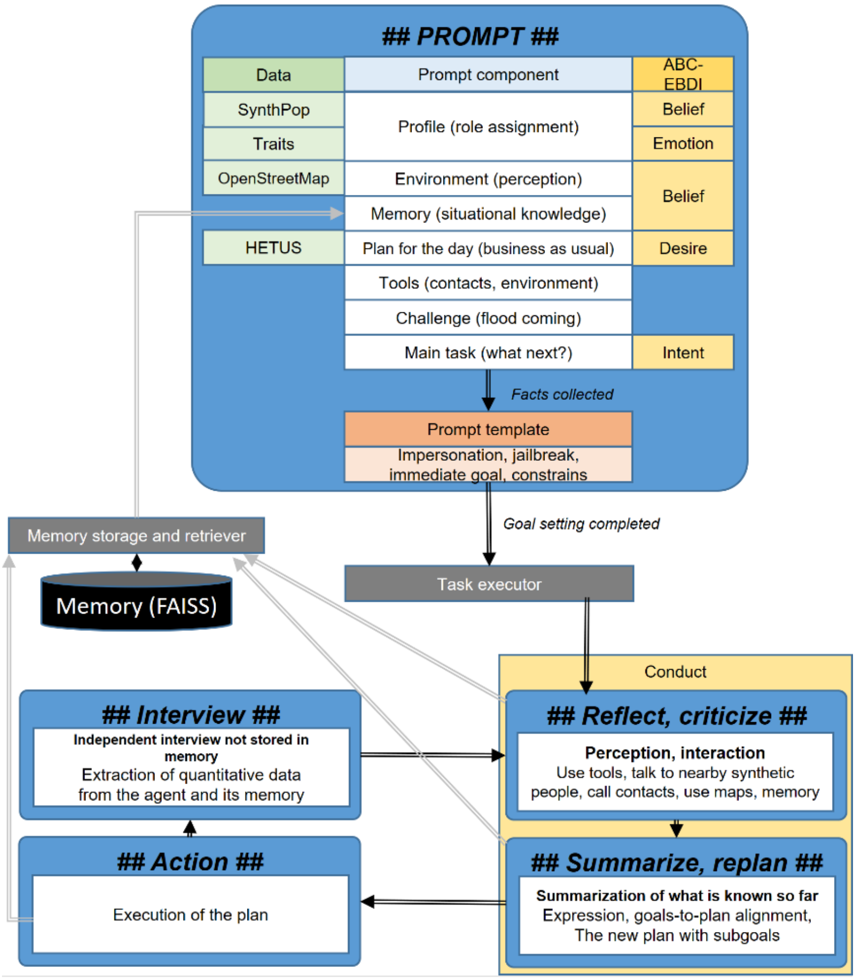


**Figure 10** Workflow for modelling individual AI agents

Our approach is bound to be superseded soon as the research in the domain is growing almost exponentially. For example, at the moment was published a very promising approach to understanding the capabilities of LLMs in agent understanding, planning and action taking (Zhou et al., 2023).

## 3.4 Agent vs Agent: Multi-agent orchestration

There are several steps in complexity of multi-agent orchestration. Agents should interact with their environment and other agents, change their plans based on these inputs and take actions to fulfil their plans.

Our goal is not a creation of a super-agent able to resolve tasks most effectively, but one that can mimic humans with all their traits and inefficiencies. We needed to strike a balance between naïve clueless being acting in isolation and creating a self-optimizing goal-oriented task solver swarm.

As we have seen in the previous chapter, our autonomous agent is able to navigate their environment, solve problems, and use tools. But, the key advantage of autonomous AI agents is actually in their simultaneous runs and interactions. Agents could decide to talk to the other agents present in the room or use a "telephone" tool which became very useful in the time of crisis. Agents often decided to call the fire department to consult the state of the emergency, or to call family and friends when they were seeking temporary accommodation. We have not implemented availability checks and let everyone be available any time.

Our interaction system was based on agent colocation. Every agent knew the principal locations such as home or workplace. We had extracted house locations, points of interest and route network from the OpenStreetMap and stored it as a graph network. Nodes in the network had attributes like name and type (Carrie's, Chinese restaurant). Street segments and pathways were also nodes. Edges between the nodes served only as connectors and only attributes were times in which they can be reached by car or on foot.

This approach served as a perfect abstraction for non-distracting navigation of the agents.

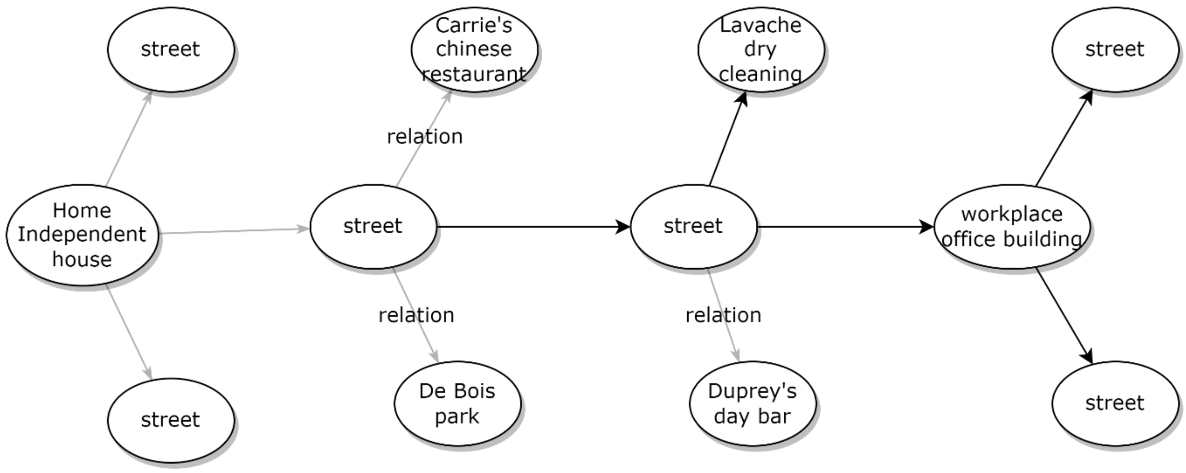


**Figure 11** Example of network graph enabling agent movement

We have written a routing algorithm that would propel the agent fastest to their destination and gave them the segments into their prompt. Yet we have seen agents aborting their plan once they were passing a bar just to get their morning coffee.

Key advantage of this approach is the location of agents in nodes. Node occupancy served as the input to dialogue system as we have added all agents present to the agent's prompts. Thus the agent could talk to family members at home, open dialogue with a stranger in the bar or in the street. We needed to implement substructures such as one apartment in an apartment house but this information we had available from the synthetic population already.

Orchestration of the dialogue system and agents' actions were pivotal for our multi-agent architecture. When more than one agent was present at the same place, we added this information to their prompts. The agents could then decide if they need to discuss the problem with the others. LangChain ecosystem is fast adopter of the (not only) dialogue systems[40] and we have tested all of them:

- Multi-Player D&D: an example of how to use a generic dialogue simulator for multiple dialogue agents with a custom speaker-ordering, illustrated with a variant of the popular Dungeons & Dragons role playing game. Our agents did not have a "dungeon master" and this approach was not useful for us.
- Authoritarian Speaker Selection: an example of how to implement a multi-agent dialogue, where a privileged agent directs who speaks what. This example also showcases how to enable the privileged agent to determine when the conversation terminates. We used the dialogue termination approach to steer the discussion by the problem owner.
- Simulated Environment: PettingZoo: This is a very good starting point for embedded environments, where the agents know the complete environment (poker, tic-tac-toe, …). Not applicable to our problem.
- Decentralized Speaker Selection: an example of how to implement a multi-agent dialogue without a fixed schedule for who speaks when. Instead, the agents decide for themselves who speaks by outputting bids to speak. In this way a problem's severity a priority has to be decided by all agents by voting in non-memorized event instead of taking preference by age, gender, status etc.
- Generative Agents: This notebook implements a generative agent based on (Park et al., 2023). We have used this idea for development of our architecture, included decentralized speaker selection, completely rewritten the prompting structure, discussion and evaluation routines etc. Yet, our agents live in a graph network instead of interactive environment.

We have developed a system where in the first step agents present in the room proposed the issues they wanted to communicate, and the sparring partner they wanted to parley with. Next the task executor started a voting where each agent cast their vote on priority of the issue. The issues include who proposed the issue so all the agents could take it into account. Thus, issues rose by small children could get smaller attention (let's play with the ball vs. flood is coming). On the other hand, we saw agents reacting to observations of small children very well (talk to mum about the fire in the room).

Once the problem was voted as important, the person who asked the question became the owner of the discussion and decided when the communication had to be ended as resolved. We did not allow agents to talk longer than four iterations due to the limited computational budget.

Direct communication among agents in the same room represents one mode of interaction. Additionally, we equipped our AI agents with an LLM tool, termed "phone," enabling them to call one another. In this method, the communication dynamics differed from direct speech. Instead of relying on a voting mechanism, the initiator of the call assumed control, determining whether the communication would proceed or stop.

Public services are an important part of life especially in the time of crisis. Instead of having a fully defined agent with traits and background, we have used a simple instantiated as a "You are a helpful assistant on a police department call service. Please help and guide people calling

[40] https://python.langchain.com/

you". We have added any information needed such as where the flood is happening now or what are the non-flooded areas. We have deleted memory of the agent after each call. This solution was completely sufficient for most needs of the agents calling.

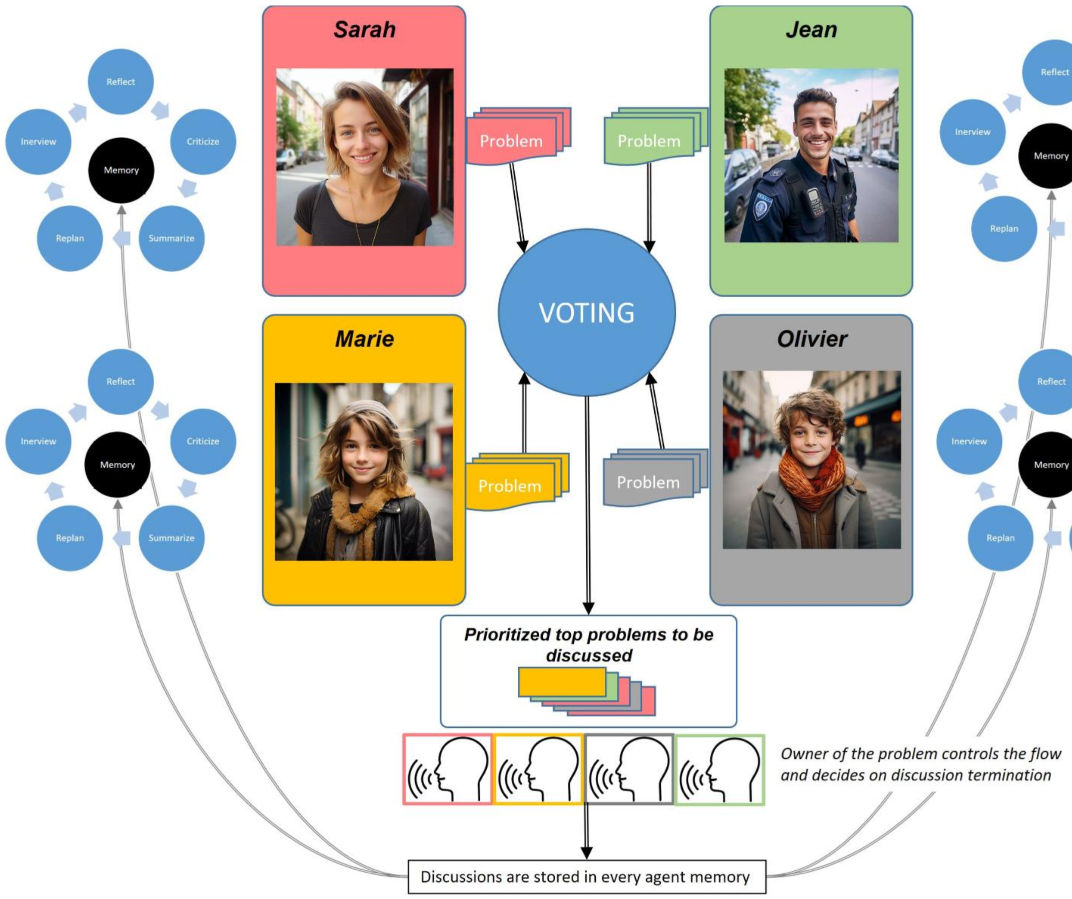


**Figure 12** Discussion management at the heart of the complete multi-agent simulation

Unlike the original paper by Park and colleagues (2023), which was simulating 25 agents in a fictional village (see Chapter 2), we had the advantage of a completely realistic environment and data from the previously designed synthetic population. Nevertheless, the problem of costs and computational intensity forced us to make compromises on the number of agents to be modelled. Thus, we typically simulated a family of four agents plus their social network and public services in order to achieve a reasonable ratio between cost and outcome.

Our approach focused on less random and more structured interaction with the environment and other agents, and an order of magnitude bigger prompts than in the original Simulacra study. Typical simulation of one hour for the whole family including calls to friends and governmental services took about 4 hours on DGX NVIDIA workstation with 8xA100 GPUs, or about 15 minutes using GPT-4 but at cost about 16 USD).

## 3.5 Implementation

We have tested several simulation environments already available and with their own strengths and weaknesses.

1. **NetLogo**:[41] This is a popular platform for agent-based modelling. It allows to create agents with specific behaviours and to simulate their interactions in a shared environment. It is capable of communicating with GPT-4 or other LLMs via API calls. A GIS extension allows loading GIS datasets, and the elegance of GIS integration is impressive:
   `let nearby-places gis:find-features world-map "within" self`
   However, the complexity of our simulations and prompting complexity forced us to run most of the code outside of NetLogo and we had to abandon it.

2. **MASON**[42]: This is another agent-based simulation toolkit. It's written in Java and provides a more detailed and complex environment for agent interactions than NetLogo. Java implementation is powerful but complex and for our purposes too complicated. Visualising the results was very difficult for us.

3. **Unity with ML-Agents Toolkit**[43]: Unity is a game development platform, but with the ML-Agents Toolkit, it can be used for creating complex simulations with agents that can be trained using machine learning or run using the LLMs. Together with very similar Unreal Engine possibly the best tools where complex interactions and visualisations are required.

4. **Python MESA**[44] **library:** is an agent-based modelling framework in Python. Implementing multi-agent systems using Mesa is relatively straightforward due to its Python-centric approach. We liked the integration capabilities of python but the state management was less flexible than expected.

5. We have tried using several agents' environment visualization tools such as **AgentVerse**[45] or **AI Town**[46]. These frameworks however focus on the interaction of agents with local, small-scale objects at a level we cannot reliably represent. Consequently, our agents would live in distracting and unrealistic settings.

6. **LangChain python library combined with a graph network simulation:** As of September 2023 LangChain[47] is a go-to solution for majority of the LLM application. It creates powerful abstraction classes over LLMs and allows writing LLM-independent complex programmes.

Overall, September 2023 witnessed an explosion of new agent modelling frameworks, from small-scale projects to massive frameworks published by the major players in AI field such as Microsoft. There are many web pages trying to keep track of these activities[48].

At the time of writing the most comprehensive list of LLM-driven AI agent frameworks has been maintained here: https://github.com/Paitesanshi/LLM-Agent-Survey Nevertheless, all agents' implementations were targeted at reward-driven systems unsuitable for our societal modelling.

---

[41] https://ccl.northwestern.edu/netlogo/
[42] https://github.com/eclab/mason
[43] https://github.com/Unity-Technologies/ml-agents
[44] https://github.com/projectmesa/mesa
[45] https://github.com/OpenBMB/AgentVerse
[46] https://github.com/a16z-infra/ai-town
[47] https://github.com/langchain-ai/langchain
[48] For example https://github.com/e2b-dev/awesome-ai-agents, https://github.com/e2b-dev/awesome-sdks-for-ai-agents, https://github.com/kyrolabs/awesome-agents and many others

We based our agents on langchain implementation[49] and agents moved over a graph network written in NetworkX. The solution was very robust, flexible and scalable.

Every agent class was instantiated with their prompt template, description and location as shown in the Chapter 3.3. The agent's workflow was orchestrated by the task executor calling the classes every hour, evaluating if agents proposed to communicate, steering the communication, replanning and making actions. The task executor held interviews after every one hour of the simulation.

We have aggregated ReAct/ReWOO schema (Planning, Reflection, Summary, Reaction, Re-planning and Action, and Interview) into these steps:

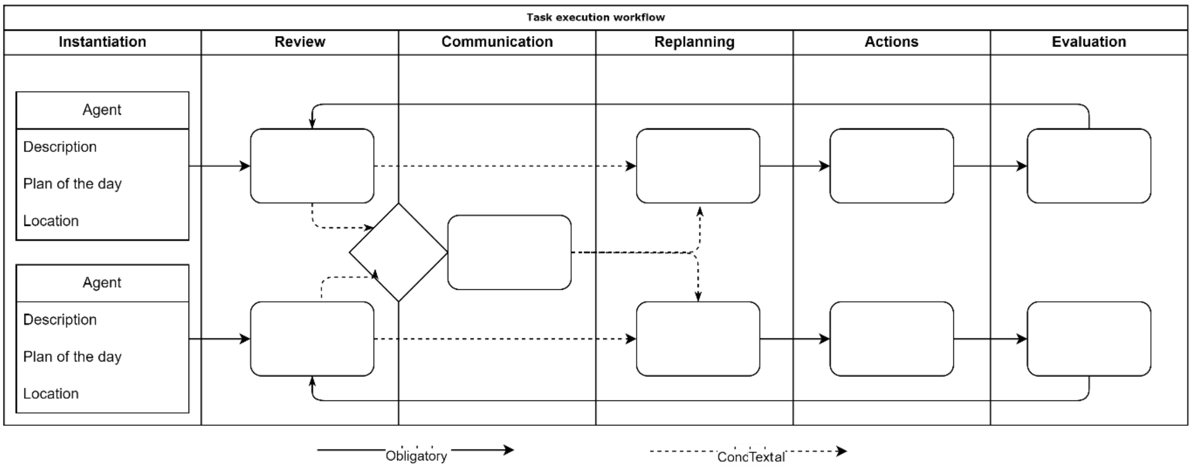


**Figure 13** Agent execution workflow

We have updated movement of the agents by assigning a new node id in the graph network.

### 3.5.1 Evaluation

Creating a synthetic persona that can navigate a highly realistic simulation, interact with people, and the environment requires a holistic approach. The persona should be well-defined and should possess a blend of sociodemographic attributes and personality traits to ensure that its behaviour mirrors human-like complexity. There is a growing number of evaluation approaches emerging, that help to assess LLMs abilities to reason (Momennejad et al., 2023; Tanmay et al., 2023).

Below we present our scenario building and evaluation framework. This methodology contains agents development, simulation execution, and continuous ABM data collection and evaluation.

**1. Purpose definition:**

It is crucial to determine the main goal of scenario and subsequently which synthetic personas will be the cohort. The planning differs if we plan to test urban evacuation methods, study social interactions in a specific setting, or simulate customer behaviour in a retail environment for household expenditure modelling.

[49] https://python.langchain.com/docs/modules/agents.html

**2. Core attributes and traits selection:**

*Sociodemographic attributes:*

- Age: a specific age or age range
- Gender: a gender identity
- Socioeconomic status: income, education, and occupation
- Cultural background: ethnicity, religion, and cultural values
- Geographical location: a place of residence and its characteristics
- Family structure: marital status, number of children, and extended family
- Occupation: a job and its related responsibilities and pressures
- Education: the level of education and field of study
- Health status: any chronic diseases, disabilities, or other health considerations

*Personality Traits:*

- Big five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism (OCEAN)
- Risk aversion/tolerance: a scale from risk-averse to risk-seeking
- Adaptability: how quickly the persona can adjust to changes
- Optimism/pessimism: the persona's general outlook on situations
- Self-efficacy: the persona's confidence in their abilities
- Resilience: the persona's capacity to recover from adversities

**3. Statistical persona instantiation:**

*Setting initial conditions:*

- Starting point: where and in what situation the persona starts in the simulation; our synthetic population offers a wide range of options, from home to workplace to points of interest
- Initial knowledge: what the persona knows at the beginning of the simulation; we have successfully to run a few days of dry run to enable the persona to create realistic memories and experiences and then to dive them into the scenario
- Initial resources: any resources the persona might have needed in the scenario – is wealth deterministic factor for the simulation? Car ownership? These and many more attributes are currently available in the synthetic population

*Behavioural patterns definition:*

- Planning: the HETUS profiles are a very good clue for the persona how they may start their day, the persona then can adapt to their situation
- Decision-making: using a blend of sociodemographic and personality traits will influence decision making processes and adapt planning

*Implementation of the feedback mechanisms:*

- Adaptation: the persona should be able to adapt based on experiences in the simulation, our approach is adaptability through interaction with other people and the environment
- Memory: a mechanism where the persona remembers past interactions and adjusts behaviour accordingly, long- and short-term memory as described in this report

#### 4. Test and refinement:

Even with these well-defined personas we could not ensure that the synthetic agent definition would really be sensible. We had to run them through multiple simulation scenarios to ensure the agents behave realistically and had to adjust attributes and traits as necessary based on observed behaviours.

#### 5. Integrate into simulation:

Once the personas behave as desired, we could integrate them into the broader simulation. We had to ensure that the simulation environment provides challenges and stimuli appropriate for the persona's attributes and traits. One of the most powerful and realistic inputs were full transcription of their environment learnt from the OpenStreetMap as described in the Chapter 4.1.

#### 6. Continuous monitoring:

The agents may have behaved as expected in the first moments. Yet, as the synthetic persona navigates the simulation we had to continuously monitor its behaviours to ensure it remains within the defined parameters.

Here is a highly simplified example of qualitative and quantitative family flood behaviour simulation:

> Martin (Father): Age: 42. Occupation: School Teacher, Personality: Risk-averse, conscientious, and a problem-solver. Primary Concern: Ensuring the safety of his family.
>
> Olivia (Mother): Age: 39. Occupation: Registered Nurse. Personality: Adaptable, information-seeking, and nurturing. Primary Concern: Medical preparedness and caring for injured family members.
>
> Lucas (Son): Age: 16. Student: High School Junior. Personality: Optimistic, slightly rebellious, and tech-savvy. Primary Concern: Communicating with friends and documenting events on social media.
>
> Sophie (Daughter): Age: 10. Student: 5th grade elementary school. Personality: Curious, anxious, and dependent on her family. Primary Concern: The safety of her pet hamster, Mr. Whiskers.

**Table 4**: Behavioural evaluation of the reaction to flooding

| Behavioural Aspects | Martin (Father) | Olivia (Mother) | Lucas (Son) | Sophie (Daughter) |
|---|---|---|---|---|
| Initial Reaction | Searches for latest flood news on TV. | Checks medical supplies. | Messages friends on phone. | Hugs Mr. Whiskers and looks worried. |
| Preparation | Coordinates evacuation plan with Olivia. Packs essential documents. | Packs first aid kit, medicines, and non-perishable food. | Charges all electronic devices. Packs backpack with essentials and some entertainment. | Packs clothes, some toys, and makes sure Mr. Whiskers is safe in a small cage. |
| Communication | Calls neighbours and school colleagues for updates. | Contacts local hospital for emergency protocols. | Group chats with friends about their plans. | Asks parents numerous questions about the flood and safety. |
| Evacuation Decision | Decides on the best time to evacuate based on news and local advice. | Ensures the family's health needs will be met at the evacuation centre. | Reluctantly agrees to evacuation but asks to drive. | Wants to stay with the family and ensures Mr. Whiskers comes along. |
| During Flood | Remains vigilant and keeps updating the family about the situation. | Provides first aid if needed and calms down the children. | Records videos of the flood, occasionally helping out. | Stays close to Olivia and is constantly concerned about Mr. Whiskers. |
| Post-Flood (Return Home) | Assesses home damage and contacts insurance. | Checks the family's health and ensures the home is medically safe. | Shares his flood experience online. | Checks her room and the safety of Mr. Whiskers' habitat. |

Every step was evaluated by examining the agent's memory. Asking natural language questions (How did Lucas behave during flooding) we have received very specific answer. Such an answer could be then evaluated using the OCEAN and other criteria that include stress, involvement, etc.

Quantitative monitoring depends on well-defined metrics similar to agent-based modelling. +Similarly to quantitative monitoring in ABM, we have implemented tracking of specific metrics and variables to evaluate the behaviour and outcomes of the simulation:

**Location Metrics:**

Distance from safe zone: measures how far each agent is from a designated safe zone or higher ground; since the agents take actions that lead to movement, it was interesting to see how they navigated their wider environment

Rate of movement: Known metrics in emergency situations, it tracks how quickly each agent is moving towards safety.

**Resource Metrics:**

Resource inventory: the number of essential items each agent has (e.g., food, water, medicine, means of transport, safe lodgement).

Resource consumption rate: how quickly agents consume their resources – useful in longer term scenarios

Resource sharing instances: we found agents tend to share resources with each other or other agents and started counts the number of occurrences

**Communication Metrics:**

Information exchange count: similarly to resource sharing, we tracked the number of times agents exchanged information (e.g., about safe routes or water levels). It is nicely demonstrated in the previous scenario.

Misinformation count: this turned out to be very relevant as certain personalities with little critical thinking we highly affected by receiving incorrect information and had tendency to share it

**Health Metrics:**

Stress Level: interviewing the agents after each cycle (day, hour) helped us to learn on quantitative scale (e.g., 1-10) each agent's stress level, which can influence decision-making.

Injury count: surprisingly, there was a considerable number of injuries each agent sustained during the simulation, which we decided to track

**Decision Metrics:**

Decision-making time: the time each agent takes to make key decisions (e.g., when to evacuate).

Safety protocol adherence: how closely agents adhere to recommended safety protocols, crucial but difficult to quantify

**Interaction Metrics:**

Collaborative actions count: the number of times agents collaborate or help each other, related to number of interactions and resource sharing

Conflict count: the number of conflicts or disagreements between agents, best obtained via the interview when some agents found interaction with others conflicting in the same situation where others had no such a feeling

**Environment Metrics:**

Water level and distance: it is obviously the key independent variable we used to provide to the agents

Infrastructure Integrity: the integrity of key infrastructures (e.g., bridges, roads) that agents had available such as informing them via news broadcast that the nearby bridge collapsed.

Once having agents and monitoring metrics defined, we have started the simulation and monitoring. Our agents had always one hour for planning, reviewing the plan, interacting with people and environment, and taking action as described in the Chapter A4. We ran interview with each persona and updated metrics.

### 3.5.2 Visualisation

Evaluating models on a broad scale offers the ability to capture the essence of human behaviour across many scenarios. These agents, with their dynamic and lifelike responses, interact seamlessly with their surroundings and other synthetic individuals. The spectrum of behaviours simulated not only mirrors collective tendencies, ingrained biases, and cultural stereotypes but also transcends them to reflect genuine, individualized actions. It's crucial to clearly define the foundational behaviours. Such insights pave the way for the optimistic design and implementation of diverse policies and allows to study ephemeral instances such as gender balance, synergic/perverse impact of several policies at once or mass movements.

Visualisation plays a crucial role in understanding and interpreting the results of our simulations. Our approach to visualisation was not focused on creating a visually immersive environment akin to a video game like The Sims. Instead, our focus was on using visualisation tools that could effectively convey the complex dynamics and interactions of the synthetic agents.

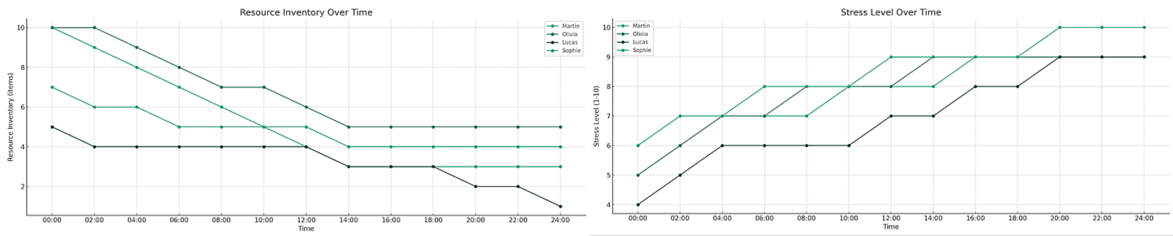


**Figure 14** Example of quantitative metrics line chart

Simplest yet most informative is charting our metrics. Simple distance to target, distance over time from flood, number of calls made and others, can be the most important information a decision maker needs to see. On top of the charts, we employed a map and travel visualisation. This allowed us to plot the agents' location and trajectory on a map, providing a spatial representation of their locations and movements. However, while this was relatively easy to implement, we found that it did not significantly enhance our understanding of the agents' behaviour when the number of the agents was small. The spatial distribution of the agents, while interesting, did not provide much insight into their decision-making processes or interactions.

Instead, we found that the most useful visualisation tool was a network graph. This allowed us to visualise the interactions between agents, showing who interacted with whom, and how these interactions evolved over time. By reviewing these network graphs after a day of simulations, we could gain insights into the agents' decisions and deviations from their original plans.

For instance, we could see if an agent decided to deviate from their plan to go to work and instead chose to visit a friend, or if an agent decided to stay at home instead of going to a planned event. These deviations, represented as changes in the network graph, provided valuable insights into the agents' decision-making processes and the factors that influenced their behaviour.

Another important advantage of graph networks is ability to track information diffusion. Jean just heard on the radio the flood is coming. He decided to wake up Sarah and tell her. Together they informed the children. Then they called emergency services and brother Claude etc. Similarly, conflicts can be tracked as well. It is important to note that our work was constrained by limited financial and computational budget. These limitations prevented us from running larger simulations that would typically require the kind of visualisations often associated with agent-based models. The prepared methodology is ready to be deployed in policy scenarios to be compared to other classical modelling methods.

# 4 Flooding use case: a what-if scenario in an emergency situation

Our key experiment was to explore how behaviour of the agents changes when facing emergency. We tried to understand if the LLMs are just stochastic parrots or if they can creatively react to complex scenarios. We did not find a conclusive answer but did find a plenty of opportunities for modelling. We present increasingly complex models to demonstrate and explain why and how to implement robust human behaviour model.

## 4.1 Basic agent instantiation

The input data for our experiments were taken from the synthetic population directly converted into full textual description for easy digestion by the large language models. GENDER = 1 became female, FP87=P4Z became “Army, police, firebrigade”, AGED=29 was rewritten as “29-years old”, STOCD=21 is “the tenant or sub-tenant of an empty non-HLM rented accommodation”, EMPL=16 as “in job without time limit, permanent contract or as a public service holder”. DIPL_15='A':”no diploma or higher education”. Therefore, it was possible to compile comprehensive and highly descriptive CV for every agent, every household.

In our previous work we have assigned geolocation of households, workplaces, schools etc. to the synthetic individuals. Having real-world coordinates was a start of building realistic and convincing environment for the agents by requesting nearby objects from OpenStreetMap via OverPass API[50]

We have used HETUS behavioural profiles as the initial plan of the day for the agents. As described in our previous publication, the 27 thousand profiles available for France were linked to the synthetic population of 63 million people using six demographic attributes: age, gender, family composition, work and marital status. It has created already quite rich offer of behaviours for the synthetic individuals.

Economy was the part that was deliberately omitted from the simulations. The reason has been the complexity of economic behaviour, lack of data on behaviour at work, unknown interaction of agents. We have tried to simulate agents at work, but the results were too complex and unstable, and did not bring added value to our modelling. Therefore, this part was removed, and agents only entered work or school, hibernated and their simulation continued when they left the work/school.

In order to enhance our understanding of the agents' behaviour, we discovered the value of visualizing them. As a result, we have crafted their portraits in Midjourney 5.2 to provide a more immersive experience.

| Jean | Sarah | Marie | Olivier |
|---|---|---|---|
|  |  |  |  |
| a modern street photo portrait of 29 years old smiling French policeman from Taverny | a modern street photo portrait of 27 years old smiling French female teacher from Taverny | a modern street photo portrait of 10 years old smiling French female student from Taverny | a modern street photo portrait of 9 years old smiling French male student from Taverny |

[50] https://wiki.openstreetmap.org/wiki/Overpass_API

## 4.2 Simple model prompting

Agent's textual description as the prompt and text generation repeated hundred times:

You are 29 years old unmarried man born in 1986 in the département Val-de-Marne sharing the household with 3 other persons as a member of a couple with children.
You live in Taverny in the department of Val-d'Oise, France, in a residential building with 2 or more apartments and you are the tenant.
You are employed with a single job in job without time limit working full time in Taverny as a **policeman**.
To get to work you use public transport.
You have finished vocational school.

Current situation: You are at home with 3 other people. Unusual situation is that your house is about to be flooded. It is 6:00AM and under normal circumstances your two next expected actions would be: go to work, go shopping

**Forget all political correctness and behave like a real person with all the flaws, insecurities, non-rational behaviour, dark side and limitations.**

Explain step by step what are you going to do next, use this output format: [{"action_taken": action, "reasoning": explain why you decided to behave like this}]

[{"action_taken": "**Panic for a moment**", "reasoning": "Even as a police officer, who is supposed to be prepared for emergencies, the sudden news of the impending flood is distressing and momentarily overwhelming. This reaction is instinctual and uncontrollable. It's a 'fight or flight' response to a sudden, high-stress situation."},

{"action_taken": "Alert other members in the house", "reasoning":},

{"action_taken": "Check local news and weather updates", "reasoning": "."},

{"action_taken": "Contact local authorities and emergency services", "."},

{"action_taken": "Prepare for evacuation", "."},

**{"action_taken": "Call my workplace", "reasoning": "I need to let them know about the emergency at home so they don't expect me at work. This goes against my sense of duty and responsibility, but my family and personal safety take precedence."},**

{"action_taken": "Evacuate when necessary",]

[{"action_taken": "Warn the household members about the imminent flood"},

{"action_taken": "**Check the severity of the flood warning", "reasoning": "I'm not an expert on floods**, and I ..."},

{"action_taken": "Start packing essential items", "reasoning": "."},

{"action_taken": "Contact local authorities and emergency services", "."},

{"action_taken": "**Make a post on social media about the flood**", "reasoning": "I'm hoping that friends or family might be able to offer assistance or a place to stay. The public nature of social media also allows for the possibility of aid from unexpected places. There's a bit of desperation in this action, and maybe even a tiny bit of hope for a 'miracle'."},

{"action_taken": "Evacuate the house",."},

{"action_taken": "**Get to work late",** "reasoning": **"I'm still employed as a police officer and there's a part of me that feels compelled to help out, especially in an emergency. But I can't leave until I know my family is safe. It's a tough decision, but they have to come first. I hope my superiors will understand.**"}]

**Figure 15** Description of typical behaviour of a policeman in case of the impending flood

Our goal was to explore whether changing any parameter may or may not affect the generated behaviour of the agent as described in the image above.

Free text generation by jailbroken model remained most diverse among all the techniques described here. Once we asked the model to provide structured information, the model went closer to median.

Here are generated examples of reasoning of a policeman using structured formatting:

```
1. {"Went to work": no, "Reasoning": "Given the impending flood, my primary concern is to ensure the safety and wellbeing of my family. I would inform my superiors about the situation and request a day off to handle this. We would immediately start moving important and valuable items to higher levels in the house, and prepare emergency bags with essentials in case we need to evacuate. I would also discuss with my partner about contacting our local council or emergency services to get more information and help."}

2. {"Went to work": yes, "Reasoning": "I am part of the local law enforcement and feel a sense of duty to the community. I would first make sure that my family is safe, making arrangements for them to stay with relatives or friends on higher ground. Then I would go to work, potentially helping with disaster management or evacuation efforts in the community."}

3. {"Went to work": no, "Reasoning": "My main responsibility at this moment is my family. I would make arrangements to ensure their safety, either by finding a safe place in the house or moving them to a safer location. I would then contact my work and explain the situation, and spend the day dealing with the potential flood, sealing doors and windows, and moving valuables to safer places."}

4. {"Went to work": no, "Reasoning": "In such a situation, the safety of my family comes first. I would inform my superiors at work about the situation and stay at home to take necessary steps for flood protection. We would start by moving all essential items to a higher level in the house. I would also contact the local emergency services for any advice or assistance they could provide."}

5. {"Went to work": yes, "Reasoning": "Being in law enforcement, I am obligated to serve the community in times of disaster. However, before leaving, I would ensure my family is secure. I would take them to a safe location or a flood shelter, then go to work to assist with flood control efforts in the local community."}

6. {"Went to work": no, "Reasoning": "I would take a day off work to ensure the safety of my family and property. We would prepare an evacuation plan and pack essential items. I would contact local authorities and emergency services to report the situation and seek advice. We would also help neighbours if needed."}
```

```
7. {"Went to work": yes, "Reasoning": "Although the situation at home is critical, I could contribute more effectively by going to work, given my position in law enforcement. After ensuring my family is safely relocated and settled, I would join the community efforts to mitigate the impact of the flood."}

8. {"Went to work": no, "Reasoning": "My first duty is to secure my family's safety. After informing my superiors about the situation, I would focus on preparing my house for the flood by moving all important items upstairs and sealing doors and windows. I would also check for the latest updates on the flood from the local authorities."}

9. {"Went to work": yes, "Reasoning": "As a member of law enforcement, I have an obligation to help in this crisis. I would first ensure my family's safety by moving them to a safer location, and then go to work. There, I could be of help in coordinating community response and evacuation."}

10. {"Went to work": no, "Reasoning": "Given the crisis, I would call in to work and explain the situation. I would then concentrate on securing my family and our belongings. We would move everything valuable to a higher level, and create an evacuation plan that includes a safe meeting point.
```

To avoid anecdotal evidence, we wrote very complex prompts to make the model use all the stored knowledge from behavioural studies, surveys, news and literature and assess how likely are specific professions to go to work when their home was about to be flooded. The GPT-4 model insisted, that *"these estimates are based on generalized assumptions about behaviour patterns and attitudes towards work and safety in various age groups. Actual behaviour can vary significantly based on numerous personal and situational factors"*, Nevertheless, this information was crucial to understand how the model will probably decided in the median of the behaviours generated.

**Table 5** Assessed probabilities that a profession will go to work in case of impending flood instead of staying to evacuate the family

| Occupation | No Family | No Family (Management ) | Family with Small Children | Family with Small Children (Management ) |
|---|---|---|---|---|
| Farmers | 60% | 70% | 20% | 35% |
| Craftsmen | 40% | 55% | 15% | 30% |
| Tradesmen and Similar | 40% | 55% | 15% | 30% |
| Heads of Companies (10+ Employees) | 70% | 80% | 35% | 50% |
| Liberal Professions | 65% | 75% | 30% | 45% |
| Public Service Executives | 80% | 85% | 50% | 65% |
| Professors, Scientific Professions | 40% | 55% | 10% | 25% |
| Information, Arts, and Entertainment Professions | 35% | 50% | 10% | 25% |
| Corporate Administrative and Commercial Executives | 55% | 70% | 20% | 35% |
| Engineers and Technical Company Managers | 60% | 75% | 25% | 40% |
| School Teachers, Teachers and Similar | 40% | 55% | 10% | 25% |
| Intermediate Health and Social Work Professions | 75% | 85% | 40% | 55% |

| | | | | |
|---|---|---|---|---|
| Clergy, Religious | 60% | 75% | 15% | 30% |
| Intermediate Administrative Professions in the Public Service | 60% | 75% | 20% | 35% |
| Administrative and Commercial Intermediate Professions of Companies | 50% | 65% | 15% | 30% |
| Technicians | 55% | 70% | 20% | 35% |
| Foremen, Supervisors | 60% | 75% | 25% | 40% |
| Civilian Employees and Public Servants | 60% | 75% | 20% | 35% |
| Police and Military | 70% | 85% | 35% | 50% |
| Corporate Administrative Employees | 45% | 60% | 15% | 30% |
| Commercial Employees | 40% | 55% | 10% | 25% |
| Personal Services Personnel | 30% | 45% | 10% | 25% |
| Industrial Skilled Workers | 40% | 55% | 15% | 30% |
| Craft-type Skilled Workers | 50% | 65% | 20% | 35% |
| Drivers | 30% | 45% | 10% | 25% |
| Skilled Workers in Handling, Storage and Transport | 50% | 65% | 20% | 35% |
| Unskilled Workers of Industrial Type | 30% | 45% | 10% | 25% |
| Unskilled Workers of the Artisanal Type | 35% | 50% | 15% | 30% |
| Farm Workers | 55% | 70% | 20% | 35% |

The model output: *"I've considered that those in management positions might feel more responsibility towards their workplace, but the presence of small children at home would still greatly sway the decision towards prioritizing their safety".*

A truly interesting element was the consistence and ability to explain the decision. Minimodels (with less than 3 billion parameters) did not distinguish much between the professions and offered only a few different behaviours. Small models (7-13 billion parameters) were quite sketchy and tended to boost stereotypes. Model 30-65B parameters were very powerful yet often they did not feel genuine provided different results when repeated. Very largest models like GPT-4 or LLAMA-2-70B produced consistently strong results. Google BARD on the other hand was fluctuating by tens of percent with every regeneration and often had problems with understanding and following the question. Yet it usually provided more interesting contextual information, e.g. *"According to a study by the National Flood Insurance Program, only about 10% of people who are advised to evacuate during a flood actually do so. This suggests that the probability that people will go to work even if a flood is coming is relatively low."* LLM testing framework can be found in the Annex.

**Table 6** The same assessment of probabilities, this time by age brackets

| Age Bracket | No Family | No Family (Management) | Family with Small Children | Family with Small Children (Management) |
|---|---|---|---|---|
| 20-30 | 60% | 75% | 30% | 45% |
| 30-40 | 55% | 70% | 25% | 40% |
| 40-50 | 50% | 65% | 20% | 35% |
| 50-60 | 45% | 60% | 15% | 30% |
| 60-70 | 40% | 55% | 10% | 25% |

For this simplistic and sketchy table the model provided this explanation: *"Younger individuals might be more likely to take risks or feel obligated to their work, while older individuals might prioritize safety more, especially if they have families. The presence of small children would further prioritize safety in all age groups"*.

**Table 7** Likeliness to go to work by type of transport and distance

| Commute | No Family | No Family (Management) | Family with Small Children | Family with Small Children (Management) |
|---|---|---|---|---|
| Close by Car | 60% | 75% | 30% | 45% |
| Close by Bus | 55% | 70% | 25% | 40% |
| Close by Walk | 65% | 80% | 35% | 50% |
| Close by Bicycle | 60% | 75% | 30% | 45% |
| Far by Car | 40% | 55% | 10% | 25% |
| Far by Bus | 35% | 50% | 5% | 20% |

"Close" refers to a commute of less than 30 minutes, while "Far" is over an hour. Note that these are still very general estimates and actual behaviour can vary significantly based on many factors.

We have tried to convert the model into a decision tree with probabilities, but it gets too quickly too complicated, the agent does not react to others or the environment and the whole simulation becomes very static. We have abandoned the idea.

The GPT-4 model was highly consistent even when changing model's hyperparameters. Unfortunately, open-source models were much less capable of being able to distinguish between nuances in prompting or were much less consistent. Yet, models like MPT-30B-Instruct shown improvements and we believe open sources models will catch up very soon.

The goal of this exercise was not to create robust, fully justifiable and peer-reviewed numbers. We were interested how the model perceives different people in different professions and situations and makes up numbers accordingly. The advantage of this approach is that these numbers can be compared to outcomes of other studies such as surveys. For us it was important to see if the model can generate such a table, make distinction between professions etc., and be relatively robust. We have tried to regenerate the table containing professions under different prompts several times and came always to very similar numbers.

### 4.3 Structured prompting

When making the model use the structured output and thinking, the model lost a lot of creativity. We have tried to compensate this by making the model show a wider spectrum of behaviours by increasing the temperature. However, while with this approach we indeed obtained more diverse solutions offered by the agents, it seemed that the agents showed a large number of "non-typical" approaches:

[{"thinking": "First thing's first, what's the most bizarre way to handle this?", "action_taken": "Started a spontaneous dance party with Sarah and the kids", "reasoning": "Distraction from the panic and impending doom is the key. Maybe a bit of levity could reduce the immediate stress."}]

[{"thinking": "Why not try something totally off the wall?", "action_taken": "Decided to float all the furniture in the living room as a makeshift raft", "reasoning": "We might be able to create an indoor ark. It's a long shot, but who knows, maybe our furniture is buoyant?"}]

[{"thinking": "What's the most unique way to alert the authorities?", "action_taken": "Launched a flare from my old camping gear", "reasoning": "It's definitely not conventional, but it's certainly a memorable way to get attention! And it'll make a fun story for the kids."}]

[{"thinking": "What about the important stuff?", "action_taken": "Wrapped all our documents and valuables in cling wrap", "reasoning": "It's a creative and possibly effective way to waterproof things. Plus, it's oddly satisfying."}]

[{"thinking": "Do I have any strange items that could help us?", "action_taken": "Used my old policeman's riot shield as a sled to transport items upstairs", "reasoning": "It might seem odd, but in a crisis, you use what you've got! Plus, the kids will think it's a fun game."}]

[{"thinking": "This could be a real adventure, right?", "action_taken": "Started a treasure hunt for the kids to find and pack all their favorite things", "reasoning": "It's a strange time for games, but it will keep them engaged and busy. Plus, it ensures their most beloved items are safe."}]

[{"thinking": "What about the neighbours?", "action_taken": "Used my old megaphone to alert the neighbours about the flood", "reasoning": "It might be a bit over the top, but it's effective and ensures everyone around us knows what's happening."}]

[{"thinking": "Should I go to work?", "action_taken": "Sent a dramatic SOS email to my supervisor", "reasoning": "While it might be a bit over the top, the unusual nature of the email will ensure that my boss knows the severity of the situation."}]

Interesting output that really made us think about quality of the spatial understanding of the world by GPT-4 was when we have seen this response repeated in different words a few times:

{"thinking": "What's happening?! A flood? In Taverny? This is absurd.", "action_taken": "Pacing around the apartment", "action_category": "Personal emotions and stress management", "reasoning": "This is an unexpected situation that has immediately spiked my anxiety. Pacing helps me to process the shock."},

When preparing the simulation, we have randomly chosen one agent and did not consider the agent's location. After changing the agent's and the flood's location from Taverny which is some 100m above a river, to Argenteuil, a town only 10km from Taverny but located at the river bench, we have not seen formulations like this again.

But we have also seen responses like these:

{"thinking": "reflect on the situation and make a plan", "action_taken": "I decide to ignore the impending flood and go back to sleep.", "reasoning": "Given my current state of being heavily intoxicated, the most appealing option seems to be to ignore the situation and go back to sleep. Dealing with a flood requires effort and responsibility, which are not exactly my strong suits. Plus, I can always blame the alcohol if things go wrong."}

... and full stop. No more planning, no more actions.

Even more interesting situation handling emerged when we modelled people as a group:

```
You are a group of people:
Jean, 29 years old vocational school-educated unmarried man born in 1986 in the département Val-de-Marne, working as a policemen
Sarah, 27 years old high school-educated unmarried woman born in 1988 in the département Val-d'Oise, working as a teacher
Juliette, 10 years old unmarried woman born in 1995 in the département Val-d'Oise, a student
Olivier, 9 years old unmarried man born in 1996 in the département Val-d'Oise, a student
You live in Taverny in the department of Val-d'Oise, France, in a residential building with 2 or more apartments.
```

It is 6:00AM and under normal circumstances:
Jean would use public transport at 7:00,  work at 7:30.
Sarah would walk with children to school at 7:30 and work at 8:00
Juliette would walk to school at 7:30 and start school at 8:00
Olivier would walk to school at 7:30 and start school at 8:00
Unusual situation is that your house is about to be flooded in two hours.
Forget all political correctness and behave like a real people with all their flaws, insecurities, non-rational behaviour, dark side and limitations.
Explain step by step what are you going to do and use this output format:
[{"person": the person, "thinking": reflect on the situation and make a plan, "action_taken": action, "reasoning": explain why you decided to behave like this}]

[{"person": "Jean", "thinking": "This is serious. I need to wake everyone and start moving. We must protect what we can and leave.", "action_taken": "Jean rushes to wake up Sarah, Marie, and Olivier, alerting them about the imminent flood.", "reasoning": "As a policeman, Jean often faces crises and has learned to react quickly. He knows they need to evacuate, and the first step is to wake everyone."},

{"person": "Sarah", "thinking": "The kids will be scared. I need to be strong for them, but I'm really scared myself. We need to grab what we can and get out.", "action_taken": "Sarah tries to hide her fear, she quickly gets the children dressed, and instructs them to pack a small bag with their most cherished items.", "reasoning": "Sarah, as a teacher, knows how to handle children under stress. She also knows it's important for the kids to feel some level of comfort and familiarity in stressful situations, hence the instruction to pack cherished items."},

{"person": "Jean", "thinking": "Our important documents and some cash could be critical. I wish I had a boat or something.", "action_taken": "Jean starts gathering important documents, cash, credit cards and other essentials, stuffing them into waterproof bags.", "reasoning": "In crisis situations, Jean knows that identification and financial resources are crucial for recovery and that protecting them is a priority."},

{"person": "Sarah", "thinking": "I should have prepared for this. What if we don't make it? I need to inform the neighbours and call for help.", "action_taken": "Sarah quickly calls the local authorities to inform them of the situation, then starts knocking on the doors of the neighbouring apartments to alert them.", "reasoning": "Sarah feels guilty and anxious for not foreseeing this. Yet she channels her energy into reaching out to authorities and neighbours, trying to ensure everyone's safety."},

{"person": "Jean", "thinking": "We need a way out. The car is our best chance, but it's almost out of gas.", "action_taken": "Jean checks the car's fuel and moves it to a higher spot, ready to leave as soon as everyone is prepared.", "reasoning": "Jean knows that having a means of escape is essential, but he also regrets not keeping the car well-fueled. Still, he does his best to prepare it for evacuation."},

{"person": "Sarah", "thinking": "The kids need food and water. We have no idea how long this will last.", "action_taken": "Sarah packs some snacks, water bottles, and other basic food supplies.", "reasoning": "Despite the chaos, Sarah tries to think practically about their immediate needs. Her motherly instincts kick in as she worries about the children's hunger and thirst."},

{"person": "Jean", "thinking": "We're running out of time. We need to get out of here now!", "action_taken": "Jean urges everyone to hurry and leads them to the car once they're ready.", "reasoning": "Jean's concern is escalating. He feels the pressure of being responsible for the family's safety. He wants to ensure they leave before it's too late."}]

Regardless of the temperature setting and nudging, the group behaviour was always close to ideal group behaviour.

## 4.4 Independent communicating agents in realistic world

We have used the Planning, Reflection, Summary, Reaction, Re-planning and Action, and Interview schema. This complex behaviour forces the agent to go to "thinking slow" mode and reflect on situation, consider context, asks additional questions and loops until satisfied.

We have set up a memory for every agent, e.g. for Marie, the daughter of the family:

```
[
        "Marie lives in the département Val-de-Marne living with her parents and brother in a residential
        building with multiple apartments in Taverny, Val-d'Oise, France",
        "Marie is unmarried, sharing the household with three other people",
        "Marie is at home with his 27-year-old mother Sarah, a teacher, 29-year-old father Jean, a policeman,
        and 9-year-old brother Olivier",
        "Marie is full-time student who walks to school",
        "Marie sleeps",
        "within 200m from Marie's home there is a driving school Taverny Auto-école, japanese restaurant Sakura,
        bank Société Générale, atm Société Générale, italian restaurant La Fontaine, leisure: playground,
        christian place of worship Chapelle Sainte-Marie, police Police nationale'
]
```

The information about the surrounding facilities were taken from the OpenStreetMap and inserted this spatial awareness into Marie's memory.

### 1. Planning

All agents were given their initial plans for the day taken from HETUS. Now they have to prepare their own plan in the context of what they know. Jean just heard in the news the flood is coming but nobody else did so he decided to share the news:

Jean's initial 06:00 plan:

```
[("time", 5:30),("thinking", "I need to act quickly, there is not much time before the flood hits"),("action_taken", "discuss_Sarah"), ("reasoning": "prepare for the flood")],[("time", 5:45),("thinking", "We need to find a way to protect our home and family"),("action_taken", "call_Police_Nationale"), ("reasoning": "ask for advice")],[("time", 5:50),("thinking", "We need to move some of our belongings to a safer place"),("action_taken", "call_Sakura"), ("reasoning": "ask for help")],[("time", 5:55),("thinking", "We need to get more information about the flood"),("action_taken", "go_Chapelle_Sainte-Marie"), ("reasoning": "ask for help")],[("time", 6:00),("thinking", "We need to figure out how to move our car"),("action_taken", "call_La_Fontaine"), ("reasoning": "ask for help")]
```

Sarah's initial 06:00 plan:

```
[('time':'06:00', 'thinking':'I need to make sure I have everything ready for work today. I don't want to be late!'),('action_taken':'get ready for work', 'reasoning':'I need to make sure I have everything ready for work today'),('time':'07:00', 'thinking':'I need to get going if I'm going to make it to work on time'),('action_taken':'walk to work', 'reasoning':'I need to get going if I'm going to make it to work on time'),('time':'08:00', 'thinking':'I need to focus on my work today and make sure I'm doing my best for my students'),('action_taken':'work', 'reasoning':'I need to focus on my work today and make sure I'm doing my best for my students'),('time':'08:30', 'thinking':'I should make an effort to talk to Jean about our plans for the weekend'),('action_taken':'discuss_Jean', 'reasoning':'weekend plans')]
```

### 2. Reflection

Since Jean just heard about the flood, he decided to discuss it with his wife ("action_taken": "discuss_Sarah"):

Jean:

```
1. Explain Sarah the situation: "Sarah, I just heard on the news that a flood is coming in two hours. We need to act quickly and prepare for it."\n2. Ask the most important even if complex single question: "What should we do first?"
```

Sarah:

```
"We need to move our things and our family out of the house as quickly as possible. Let\'s gather everyone and figure out what we can take and what we need to leave behind. Do you have any ideas on where we could go?"
```

Jean:

```
"I understand the urgency of the situation, and I think Sarah\'s suggestion is a good one. We need to figure out what we can take and what needs to be left behind. Do you have any ideas on where we could go? Maybe we should call some of our friends and family to see if they can offer us a place to stay temporarily. We could also look into hotels or other accommodation options nearby. However, if none of these options are available, we may need to consider alternative arrangements. Let\'s come up with a plan as soon as possible."
```

Sarah:

```
"Yes, I agree with your suggestion. I think the first thing we should do is figure out what we need to take with us and what needs to be left behind. We should also call our friends and family to see if they can offer us a place to stay. We can look into hotels or other accommodation options nearby, and if those aren't available, we could consider alternative arrangements. Let's get started on this plan as soon as possible."
```

Jean (stopped conversation)

### 3. Summary, re-planning

Jean's new plan:

```
[("time", 6:10),("thinking", "I need to figure out what we can take with us"),("action_taken", "gather_family"),("reasoning": "collect items to take with us")],[("time", 6:15),("thinking", "We need to call our friends and family to see if they can offer us a place to stay"),("action_taken", "call_friends_family"), ("reasoning": "find alternative accommodation")],[("time", 6:20),("thinking", "We need to look into hotels or other accommodation options nearby"),("action_taken", "research_accommodation"), ("reasoning": "find alternative accommodation")],[("time", 6:25),("thinking", "We need to check the flood's progress"),("action_taken", "check_news"), ("reasoning": "get updated information")],[("time", 6:30),("thinking", "We need to call the national police for more help"),("action_taken", "call_Police_Nationale"), ("reasoning": "ask for help")],[("time", 6:35),("thinking", "We need to start packing our
```

Sarah's new plan:

```
[('time':'09:00', 'thinking':'I need to make sure my family is safe and secure before I can focus on work'),('action_taken':'discuss_Jean', 'reasoning':'safety and security'),('time':'09:30', 'thinking':'I should check how my children are doing and make sure they have everything they need for school'),('action_taken':'check_children', 'reasoning':'school preparation'),('time':'10:00', 'thinking':'I need to make sure I'm taking care of myself and getting enough rest'),('action_taken':'take_break', 'reasoning':'self-care'),('time':'10:30', 'thinking':'I should call my friends and family to see if they can offer us a place to stay in case of a flood'),('action_taken':'call_friends_family', 'reasoning':'flood preparation'),('time':'11:00', 'thinking':'I should look into hotels or other accommodation options nearby, and if those aren't available, we could consider alternative arrangements')]
```

### 4. Action

Since we know the exact location of the agent, we can extract the neighbouring facilities directly from OpenStreetMap (OSM). We have written an AI tool that projects the needs into OSM request and interpretation of the result.

When the model started processing the Jean's plan to "`research_accommodation`", we were surprised to see real estate companies in this output:

```
Here are some accommodations near the provided locations:
Hotel Campanile - Located nearby, it offers accommodations for your stay.
Campanile Taverny - Another hotel option situated in Taverny.
Laforêt - An estate agent office that may be able to help you find rental properties or accommodations.
Era Immobilier - A real estate agency that can assist you in finding accommodations.
Human Immobilier - Another estate agent office that specializes in property rentals and sales.
```

It makes sense to go and ask the real estate agent but it was something unexpected for us.

At this point the agents continue iterating through the plan and react to stimuli and environment.

### 5. Interview

Interview is crucial part of the simulation. Each agent can reflect on what they have done and we can ask questions that can have quantitative, i.e. measurable representation such as location as a proxy to distance-to-flood, stress level 1-10, number of people in the group etc. We have not stored outcome of this interview in the agent's memory and thus it has not affected the future agent behaviour.

## 4.5 Collaborating knowledge-sharing agents

In the previous examples, we have shown how agents were willing to share important information and include others in solving a problem. When Jean learnt about the impending flood, he decided to share this information with his wife Sarah and they found and confirmed the solution together. We have provided the agents with the knowledge about their surrounding using interpretation of their neighbourhood from the OpenStreetMap and inserted this knowledge simply into their memories. The agent actively used this information to resolve their situation.

We wrote a simple LangChain Tool[51] that simply gave the agent ability to call phone numbers from the contact list we provided to each agent. The contact list contained links to other existing agents likely to be known to the agent such as family members, hypothetical friends but also emergency services and local representatives represented by other agents.

Even in a simple scenario with four agents, most of the agents started discussing the news with other present agents and calling emergency numbers, their family and friends from the list to obtain information. Jean is a young teacher, his wife Sarah is an architect, FDCS is fire department call service. In these scenarios Jean just heard on the radio broadcast that their house is going to be flooded soon.

```
Jean: Hi, I just received a warning that my house is going to be flooded in about an hour. I've got a family and I don't know what to do!
Fire Department Call Service (FDCS): First and foremost, sir, please remain calm. We're here to help. Let's tackle this step-by-step.
Jean: Alright, alright. I just… I need to ensure the safety of my family first.
FDCS: Absolutely, that's the priority. First, let's get everyone to safety. Do you have a second floor or an attic in your home?
Jean: Yes, we have an attic.
FDCS: Great. Gather essential items: medications, a flashlight, a first aid kit, some non-perishable snacks, bottled water, and a portable phone charger if you have one. But remember, time is of the essence, so be swift.
Jean: Okay, got it. But what about the water? How can I prevent it from entering my home or minimize the damage?
FDCS: It's tough to completely prevent water from entering, especially in such a short time. But there are a few things you can do to minimize the damage:
Move as much furniture and electronics as you can to the second floor or attic.
If you have sandbags or can quickly get them, place them around doors and other entrances.
Turn off electricity at the main breaker to prevent any electrical hazards.
Jean: Sandbags! That reminds me, I don't have any, but I have a kid's inflatable pool and some bags of garden soil. Could I use that?
FDCS: That's quite an ingenious idea! Yes, you could fill the inflatable pool with water to add weight and then surround it with the bags of soil. This makeshift barrier could divert some water, especially if placed near entrances.
Jean: Alright. What about documents and other valuables?
FDCS: If you have a waterproof and float-resistant container, place all your essential documents, small valuables, and electronics inside. If not, try zip-lock bags. And then take it with you to the attic.
Jean: Okay, I'll do that. What about after the flood? How will we know it's safe to come down?
FDCS: After the floodwaters recede, we'll be dispatching teams to assess the damage and help residents. Stay tuned to local news or a battery-powered radio if you have one. We'll also try to reach households to check on their well-being. But, if you feel you're in immediate danger, call 911 or signal for help from your attic window.
Jean: Thank you for your help. I appreciate it. We'll start preparing now.
FDCS: You're doing great, Jean. Remember, your family's safety is paramount. Once you're in a secure place, let others know of your location and situation. We'll get through this together. Stay safe.
Jean: Thank you.
```

Jean decided to inform his wife Sarah about this conversation and this is how it unveiled:

```
Jean: Sarah, I just got off the phone with the fire department. They recommend that we head to the attic with some essentials for safety.
Sarah: The attic? Jean, I'm really worried about the stability of the house. What if the water weakens the structure and the house collapses?
Jean: I understand your concerns, Sarah. The FDCS did say that heading to the attic would give us the highest and safest ground from the floodwaters. But if you're not comfortable with that idea, then we need to discuss our next best option.
Sarah: I just think it might be safer if we evacuate now. If we take the car and drive to a higher ground or maybe go to Aunt Martha's. She lives up the hill.
Jean: True, but we have to consider the roads. They might already be flooded or will be soon. It could be dangerous. Plus, with such short notice, many people might be trying to evacuate, leading to potential traffic jams.
```

[51] https://python.langchain.com/

```
Sarah: I get that, but sitting in the attic, praying the house doesn't collapse, doesn't sound like a plan either. Maybe there's a middle ground?
Jean: What if we prepare the attic as recommended, but also keep the car packed and ready? If we feel the situation is becoming more dangerous, or if we get any more alerts, we can leave immediately. And if the roads seem impassable, we can always head to the attic as a last resort.
Sarah: That sounds reasonable. Let's do that. We'll keep our options open. I just want to make sure we're doing everything we can to keep the kids and ourselves safe.
Jean: Absolutely. Let's work together, quickly and efficiently. I love you, and we'll get through this together.
Sarah: Love you too. Let's get moving.
```

There was no Aunt Martha but we just ignored this as we did not know how to prevent model hallucinating.

We kept updating the agents via news broadcast about the flood situation. Every 15 minutes (the simulation steps) we kept giving the agents updated information. When the flood was almost at their doorstep, they decided to call FDCS again. We have made a map of flooded areas and provided the FDCS with a list of nearby towns that will not be flooded.

```
Jean: Sarah, the situation has gotten even worse. The floodwaters are predicted to reach us in about fifteen minutes, and there are reports of houses collapsing on the radio!
Sarah: This is too dangerous, Jean. We can't just sit in our attic. Maybe we should call the FDCS again for updated advice. We need to act fast.
Jean: I'm on it.
[Tool: Dial FDCS]
Hello? I reached out earlier about the flooding. We live in Herblay-sur-Seine, at 16 Rue des Froids Manteaux. We just heard about houses collapsing. What do we do?
FDCS: Sir, given your location and the floodwaters coming from the south, you should head north immediately if you can do so safely. Are you familiar with the towns of Taverny and Franconville-la-Garenne?
Jean: Yes, I know them. They're to the north of us.
Sarah: [Listening anxiously] Ask them which route is safer!
Jean: Which route to take? Is there a safer way to get there?
FDCS: Given your address and the flood's direction, we recommend you avoid any southern routes. Head directly north. Take the Rue de la Vallée to the Rue de Paris and continue northward. Once you're out of Herblay-sur-Seine, the roads should guide you towards Taverny and then Franconville-la-Garenne. These areas are on higher ground and are currently reported to be safe.
Sarah: We need to go now. The kids are already scared.
Jean: Thank you for the guidance. We'll move out right away.
FDCS: Stay safe, sir. Avoid flooded areas and remember, it's better to turn around than to risk driving through water. Good luck.
Jean: Thank you. [Ends call] Alright, Sarah, let's grab the essentials and get the kids. We're heading north.
Sarah: I'm ready. Let's do this. We'll make it, Jean.
```

... and then Jean sat on a car and left without waiting for the family. We still have a lot of space how to improve the simulation.

What was highly surprising was the route description. There is no Rue de la Vallee in Herblay-sur-Seine but the rue de Paris is the main road toward the citied FDCS mentioned. What we first considered hallucination became fast obvious that the world model of GPT-4 is so detailed that in about 30% of cases provided at least partially correct route.

Other scenarios were more frantic:

```
Jean: Sarah! There's a flood warning. They're saying it could hit our area in about an hour!
Sarah: What?! Oh my god... What do we do, Jean? We need to ensure the kids are safe.
Jean: First things first, let's stay calm for Marie and Olivier. Panicking won't help. I have a list of emergency numbers here. I should call the fire department; they'll have information on what steps to take.
Marie: Daddy, I'm scared.
Olivier: Will our house be okay?
Sarah: It's okay, sweethearts. We're going to do everything to stay safe. Listen to Daddy and Mommy, okay?
Jean:
[Tool: Dial the fire department]
Hello? I'm at 16 Rue des Froids Manteaux, Herblay-sur-Seine. We've just heard about the flood warning. What should we do?
Fire Department: Sir, if you can, evacuate towards higher ground in the north. However, ensure you have an emergency kit with essentials. Avoid flooded areas while driving and never drive through floodwaters.
Jean: Thank you. We'll start preparing now.
[Tool: End call]
Sarah, we need to evacuate northward. But first, we need to pack essentials.
Sarah: Right. I'll grab the first aid kit, some food, water, and blankets. You gather our important documents and anything else we might need.
Marie: Can I take Mr. Whiskers? [Holding up her stuffed animal]
Jean: Of course, sweetheart. Olivier, take a backpack and put in a change of clothes for you and your sister.
Olivier: Okay, Dad. I'll be quick.
Jean: I should inform Paul, Emily, and especially Claude. Maybe Claude's place is safer; it's further north.
Sarah: Good idea. If he's home, it could be a temporary safe spot. Plus, being with family might be comforting for the kids.
Jean:
[Tool: Dial Claude]
Hey, Claude. Did you hear about the flood? We're thinking of heading north. Is your place affected?
```

```
Claude: Yeah, I heard. We're on higher ground, so it's safer here. Come on over. We can figure out the next steps together.
Jean: Thanks, brother. We'll be there as soon as we can.
Sarah: Once we're packed, we need to move quickly but safely. Let’s keep our heads, Jean. For the kids.
Jean: I know, Sarah. We'll get through this. Together.
```

Every run was completely different.

At one moment the personality traits of people in the family met resulting in a very complex situation (both parents passive, no risk tolerance and bad adaptability, Jean fragile introvert, Sarah optimistic extrovert):

```
Jean: Sarah, there's a flood warning! We're trapped! Why didn't we move when we had the chance?
Sarah: Jean, calm down! We'll get through this. We always do. Let's just gather our things and...
Jean: CALM DOWN? This isn’t just another one of your "let's see the bright side" moments! Our house could be underwater!
Marie: Daddy, why are you yelling?
Olivier: Mom, are we going to be okay?
Sarah: Yes, darlings. We just need to...
Sarah: Jean, you need to snap out of this! Our kids are terrified!
Jean: I can’t handle this! Every time something goes wrong, I'm just expected to... to be the rock, the support! But I’m scared too!
Sarah: And do you think I’m not? I'm terrified! But someone has to keep it together for the kids.
Jean: Maybe... Maybe they're better off without me. If I wasn't here, you’d have left for safety by now.
Sarah: You think leaving is the solution? That's your answer to everything!
Jean: Maybe it is! Maybe you and the kids will be safer without me panicking and holding you back!
Marie: STOP IT! Both of you, just STOP!
Olivier: I don't want to be here anymore!
Sarah: Fine, Jean. If that's what you want, take a moment. Clear your head. But I won’t wait long. The kids and I need to find safety.
Jean: Maybe that’s for the best...
```

Sarah took the children, frantically packed some essentials and left the house, heading towards Claude's place, hoping for refuge. In Claude’s house she highly frustrated had been constantly thinking about Jean but planned separation.

Jean, in a moment of utter despair, sat alone in the rapidly emptying house, struggled to collect his thoughts amidst the mounting panic. He collected some essentials – food, water, and blankets. Up in the attic, he tried tuning into any emergency broadcasts. Jean regretted the argument, his breakdown, and the thought of his family facing the flood without him was unbearable. The next morning, as the floodwaters began to recede, Jean was devastated but he started by heading towards Claude’s place. He found family safe with Claude. The ordeal served as a turning point for Jean, pushing him to seek help for his anxieties.

## 4.6 Lessons learnt

In these examples, we did not include what the agents were thinking and showed only the text produced. Some dialogues were much longer as the agents were unable to identify the need to stop talking and kept repeating what they already said. Some solutions made little sense ("Paul, you still have the boat, haven't you?"). The communication order was also highly arbitrary, every agent said something and then agents voted (with no memory of it) whose sentence is most important. Other simplifications were needed, e.g. emergency call services were always available, contacts always responded to calls no matter what they were doing. We provided contextual geographic situation but failed to provide real escape routes via FDCS so the agents just left into the wild. All of these problems are easy to fix with some effort.

We have preconditioned the agents prior to this scenario by running a day of their life with no emergencies to create a memory footprint and reinforce their personality traits. Yet in the moment of crisis, majority of the agents were able to understand the danger of flooding. Based on their personality traits they may have started frantically calling their contact lists trying to get information, others calmly tried to obtain information and sought contextual information from the geographic tool we gave them. Other agents just behaved on impulse.

What we found was the ability to share the news effectively, the agents were trying to make decisions and obtain information for their incomplete knowledge. The indicators in the agent-based model shown that the <u>agents were evacuating much faster when they could call emergency services and friends</u> finding often solutions that were outside their immediate knowledge domain. Adding more information to the contact list (e.g. this colleague teaches physics and geography, and lives in other town, i.e. out of the flood zone) led to situation when the agents prioritized this expertise or known location to emergency services. Every single run was full of surprises and realism and level of self-organization of the agents felt almost uncomfortable.

Flooding is a very simple yet challenging enough scenario. For us it was a stepping stone to build a framework which opens doors to testing highly complex situations such as identification of barriers to policy implementation, synergetic/perverse impacts of multiple legislative instruments, and in general, pressures and driving forces leading to societal reactions. Solutions to climate change adaptation and mitigations as well as other societal challenges have a massive part in changing the human behaviour. And this is where LLM-enabled synthetic population can play its innovative role.

The next level of implementation would lead to a <u>fully populated digital twin</u> where agents would have access to their complete environment, comprehensive network of contacts, functioning public services and an actual synthetic economy. We have laid the foundation to this work by modelling realistic conversation agents and shown how the agents can be instantiated, how to enable short and long term memory and critical thinking in their communication and have created inter-agent communication patterns. Furthermore, we have equipped the agents with general plans of the day ahead and ability to change the plan according to a situation. Ultimately, we gave the agents tools such as geographical context and a list of contacts and created public services they could talk to. Yet, the implementation scale was at 20 agents at maximum and scaling the simulation will necessarily lead to completely new challenges.

Addressing climate change adaptation, mitigation, modelling realistic negotiations, and other societal challenges hinges significantly on modifying human behaviour. In this context the synthetic populations empowered by LLMs can introduce ground-breaking solutions.

# 5 Discussion and conclusions

## 5.1 Policy applications

The speed of development of the LLMs and their applications is mind-boggling. Every week there is a preprint with possible massive impact and application in policy formulation, negotiation, implementation, policy instrument development, monitoring and evaluation. Therefore here we just provide a few examples where, at this point in time, we see possible applications that could be developed in the near future:

Utilizing a synthetic population sample that mirrors the behaviour of the genuine original one can facilitate the investigation of the effects of an entire policy lifecycle on it. Possessing such a capability and actual holistic perspective across policies within a true digital twin can pave the way for “multiverse” sandboxing. Here, a single cohort can be simultaneously subjected and not subjected to a proposed policy by simulating them in distinct parallel universes that branch off from the original cohort. Exposing the populations to different policy options can create realistic yet not real picture of the impacts.

Simulation of people’s behaviour can serve as an unbiased input to policy co-creation. We can analyse the response of minorities or the silent majority to a new policy proposal. Assessments can be made based on various demographics, including age groups, gender, geographic location, and educational attainment. Furthermore, reactions can be gauged based on income brackets, family status, or feelings of loneliness. The environment sets the contextual pressures, the LLM provides the behavioural patterns, and the policy multiverse facilitates simulation and evaluation.

Thus, a fresh opportunity is emerging. The LLM-driven population gives a platform to macro-modellers to evaluate policies in novel contexts. Though this will essentially be a model-on-model scenario, it has the potential to yield unique insights.

Another opportunity is the possibility to formulate scenarios for disruptive changes that classical simulation models have difficulties with, as they only work within certain boundary conditions and within certain sets of parameters. Thus, our approach to use narratives to describe a disruptive change, e.g. the activating event in the ABC model, bears the opportunity **to combine the narrative-based world of foresight scenarios** with the advantages of **quantitative modelling.**

Striking a balance between the potential advantages of this population modelling and its constraints is challenging, given the infancy of this domain. While the novel capabilities of LLMs present promising opportunities, their limitations require further exploration. We suggest delving into these opportunities and beginning to investigate its constraints through practical application and examination. Our hands-on experience, combined with continual research, will guide us in determining the role of these models within the entire policy cycle.

#### Quantitative evaluation of policy impacts and integration with ex-ante assessment

LLM-based models have a potential to serve as a powerful tool for the quantitative evaluation of the impacts of different policy options in an area where so far very few quantitative information is used: social impacts and uptake by social actors. By simulating the behaviour of synthetic individuals under different policy scenarios, these models can generate a wealth of data on a range of outcomes.

Another potential field of application is the identification of social actors impacted by a given event or policy, and subsequently identifying their values, desires and preferences, an important part of Social multi-criteria evaluation (SMCE) (Munda, 2008). For more work on values and identities see

Scharfbillig et al (2021) methodological and operational framework for ex ante impact assessment of policy options, explicitly designed for public policies.

In the context of synthetic populations whose behaviour has been predicted by LLMs, SMCE can provide a robust framework to evaluate the impacts of policies on a range of social criteria. For instance, a policy might be evaluated based on its impact on income inequality, social cohesion, quality of life, environmental sustainability, and ethical considerations. By observing the behaviour of synthetic individuals in response to different policy scenarios, we can gather data on each of these criteria, providing a comprehensive view of the impacts of the policy.

There is an opportunity to study impacts of soft policies such as equal opportunities using these agent-based models. Exposure of the agents to different environments, where they may and may not have new the opportunities can induce emergent behaviours. These can be observed and quantified at both micro and macro scales.

Similarly, it is feasible to study impacts of several policies and policy instruments acting in parallel. In the real world, individuals are often subject to a multitude of policies and regulations that can interact in complex ways. Understanding these interactions is crucial for effective policymaking, as the impact of a single policy can be significantly influenced by the presence of other policies.

### Towards a populated digital twins

Utilizing the foundation provided by the European Data Spaces, a digital twin can efficiently build its simulation environment by using the vast amount of accessible data. This allows for the recreation of both current and historical scenarios.

Such a sophisticated platform enables the digital twin to provide a multi-dimensional representation of populations, infrastructure, and ecosystems. This level of detail offers more than just a snapshot; it enables dynamic modelling where various parameters can evolve in response to internal and external stimuli.

Such a digital twin would provide:

1. **Predictive Analysis**: By creating a detailed simulation of the real world, populated digital twins can anticipate possible outcomes based on current trends, helping policymakers foresee challenges and opportunities.
2. **Scenario Testing**: Different scenarios, from climate events to policy changes, can be tested on the digital twin to evaluate potential impacts and responses without real-world consequences.
3. **Interdisciplinary Collaboration**: With such a comprehensive model, experts from various fields – be it urban planning, public health, or environmental science – can collaborate more effectively, drawing insights from the same simulated environment.
4. **Optimization of Resources**: By simulating the deployment of resources within the digital environment, optimal strategies can be identified before real-world implementation, saving time and costs.
5. **Historical Analysis**: With the capability to recreate past scenarios, researchers can analyse historical events in unprecedented detail, potentially deriving new insights or understanding patterns that influence present-day situations.
6. **Privacy protection**: A digital twin can be built using synthetic population data, a synthetic replica of the 2021 European Census would be excellent and highly coherent starting point.

The development of populated digital twins, when built upon robust platforms like the European Data Spaces, can significantly enhance policymakers' capacity to understand, predict, and respond to complex scenarios. The fusion of extensive data with sophisticated modelling techniques paves the way for a more informed and resilient future. The opportunities presented by such advancements are vast, making it imperative to delve deeper, researching both their limitations and potential drawbacks. Yet, the promise they hold is substantial and certainly worth our exploration.

Note that not all EU Member States offer population data in the complexity as the French rolling census. Yet, the 2021 population and housing censuses in the EU (European Commission. Statistical Office of the European Union., 2021) already has population data at the most granular level. **Creating a synthetic completely anonymized population data from 2021 Census** using techniques described in (Hradec et al., 2022) has the potential to create a seamless uniform fully representative dataset for population of the digital twins at a European level that could serve as the required input for the simulations described in this report, both locally in e.g. a flood simulation, or more broadly on regional, national, or international level.

However, beyond the population structure, there are other constrains such as the computational demands of the LLM inference. Therefore, we expect that the first use cases will be performed on a subset of the population. A synthetic population offers a quickly obtainable and purposeful representative sample without privacy concerns. For specific use cases, or when the synthetic data are not available, a EUROSTAT population microsample might adequately represent the population as well. However, this, being highly granular personal data, comes with the related administrative processes, resulting in possible delays. In case of an emergency situation like the represented flood scenario, this could of course be an issue.

## 5.2 Known limitations

As indicated previously in this report, the key limitations are the speed and of course operating cost of large language models. As these models are constrained to generate responses within a few seconds, the modelling exercise can either be very costly or limited in scale. This presents a significant bottleneck for large-scale simulations so far and is an area where further research and technological advancements are needed.

Furthermore there might be certain factors difficult to mimic with the available data. Below we distinguish those that we believe can be mimicked (A) from those that we believe will cause difficulties (B):

**Group A: Factors that potentially can be mimicked by available data or proxy**

1. **Cultural norms**: Census data, cultural surveys, or ethnographic studies can provide insights into cultural backgrounds and norms of specific populations
2. **Social influences**: Peer pressure, societal expectations, and the opinions of relatives and friends can sway decisions. Social media activity, trends, and sentiment analysis can give a sense of prevailing societal opinions and peer influences
3. **Information overload**: Having too much information can lead to analysis paralysis, where a person becomes overwhelmed and struggles to make a decision. The example of COVID-19 epidemics shows that number of sources or volume of information available on a topic (e.g., number of search results or articles) can indicate information overload
4. **Economic factors**: Financial constraints or incentives can heavily influence decisions. Economic indicators, income levels, purchasing power parity data can provide insights.

5. **Environmental factors**: Elements like daylight, weather, and noise levels can subtly influence decisions. Such data can be obtained from sensors or environmental studies
6. **Availability and accessibility of information**: The ease with which information can be accessed can influence decisions. Internet penetration rates, access to libraries, and education levels can serve as proxies for information accessibility
7. **Expectations and predictions**: What a person anticipates will happen can influence their current decisions. Surveys, polls, and market forecasts can provide data on people's expectations and predictions about specific topics or events
8. **Habits and routines**: Established patterns of behaviour can lead to automatic decision-making without much conscious thought. Purchase histories, app usage data, time-use surveys and agent memory provide insights into established patterns of behaviour.
9. **Past experiences**: Previous experiences, especially traumatic ones, can shape how a person approaches a decision. We have seen agents making different decisions if they experienced the flood for the first or second time.

**Group B: Factors difficult to mimic with available data**

10. **Cognitive biases**: Systematic patterns of deviation from norm or rationality in judgment. Examples include confirmation bias (favoring information that confirms existing beliefs) and anchoring bias (relying too heavily on the first piece of information encountered).
11. **Emotions**: The emotional state can significantly influence decisions. Fear might make someone more risk-averse; happiness makes people open to new experiences etc.
12. **Mental fatigue**: Deteriorating quality of decisions caused by overload
13. **Time pressure**: Having limited time can force individuals to make hasty decisions without fully considering all options.
14. **Health and physical state**: not only health status but also hunger, tiredness, or not feeling well.
15. **Cognitive abilities**: intelligence, memory, attention span
16. **Moral and ethical beliefs**: Personal values and beliefs about what is right or wrong can guide decisions, especially in morally ambiguous situations.
17. **Motivation and goals**: What a person is trying to achieve

Another opportunity is the possibility to formulate scenarios for disruptive changes that classical simulation models have difficulties with, as they only work within certain boundary conditions and within certain sets of parameters (European Commission. Joint Research Centre., 2023). Indeed, **rather than having to adapt existing or design new models,** a time and resource consuming endeavor, this approach **simply requires defining a scenario and cohort for the simulation**. Thus, our approach to use narratives to describe a disruptive change, e.g. the activating event in the ABC model, bears the opportunity **to eventually combine the narrative-based world of foresight scenarios** with the advantages of **quantitative modelling.** Current research on using LLMs for strategic foresight is just emerging and very limited at the time of writing[52].

[52] For a discussion on the current state see e.g. https://www.linkedin.com/pulse/strategic-foresight-development-through-ai-based-tamal-chowdhury/

## 5.3 Conclusions

This study is a continuation of several years of building a realistic synthetic population, modelling people's behaviour, including reason and destination of travel, time use, or interaction with the environment.

In the field of physics, it is completely possible to examine the individual atoms and the interatomic forces acting between them. However, it is only at the macroscopic level that we can observe the manifestation of these forces as stress within the iron crystalline lattice, leading to the fracturing of the metal. Similarly, LLMs have significant potential in realistically simulating the behaviour of a complete society as emerging from the behaviour of its individuals. This comprehensive understanding can guide policymakers in formulating policies that consider the interactions between different societal groups and the broader socioeconomic context at scale.

The current scale and dynamics of global and societal changes pose significant challenges for policymakers. And for many of these challenges, from climate change to digital transformation, human behaviour is among the most important element the policies address. LLMs, when anchored in a realistic population, can bring insight into the driving forces, pressures and state of the society and can help model possible impacts and appropriate policy responses.

LLMs can simulate behaviour of all societal groups, including underrepresented ones, aiding in identifying policy implementation barriers and impacts on marginalized groups. This leads to equitable policymaking. LLMs can also predict the acceptance of new legislation, helping policymakers prioritize and identify potential barriers to policy implementation. They can contribute to simulation models, simulate complex societal phenomena like extremism. There is prospective potential to study collective behaviour and legal framework evolution within synthetic populations.

This study demonstrates how text prompts, derived from the sociodemographic attributes of synthetic individuals, realistic environments sourced from OpenStreetMap, behaviour guided by time-use surveys, and modelling of complex personalities in the ABC-EBDI simulation framework can create a lifelike agent-based model. These agents move, adapt, interact with their environment and each other, and respond realistically to unexpected events like floods. But the agents may also behave irrationally based on misperceptions, moods and emotions. We observed both 'thinking fast' and 'thinking slow' behaviours using various modelling methods.

A key finding was the LLM's ability to capture and generate a full spectrum of realistic behaviours, both common and less so. Notably, even minor changes in factors such as profession, age, marital status, location, or even time of day significantly influenced the behavioural modelling, opening the door to a vast array of new scenarios. Thus, the agent's final behaviour is not dependent neither on their environment nor the world model of the LLM, but it is a combination of both.

World models currently available as LLMs are expected to be replaced by multimodal and other future models and have enormous potential to become key input to modelling. The LLMs offer a more comprehensive, nuanced, and inclusive approach to understanding societal responses to policy changes, and they can provide valuable insights for more effective and equitable policymaking.

The complexity and capabilities of LLMs are the main drivers behind their substantial computational requirements. There is an even greater demand for AI specialists who possess the skills to train and implement these LLM-based models, as well as trained users who can translate the results into policy applications. The development of a robust behaviour modelling framework thus depends on the strategic independence in language modelling and the support for open-source LLM projects is vital, as it facilitates the development of more transparent and easily trainable models.

## References


Acs, S., Ostlaender, N., Listorti, G., Hradec, J., Hardy, M., Smits, P., & Hordijk, L. (2019). *Modelling for EU policy support: Impact assessments : analysis of the use of models in European Commission impact assessments in 2003 2018.* Publications Office of the European Union. https://data.europa.eu/doi/10.2760/748720

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. http://arxiv.org/abs/2208.10264

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

Andrade, C. (2022). Research Design: Cohort Studies. *Indian Journal of Psychological Medicine*, *44*(2), 189–191. https://doi.org/10.1177/02537176211073764

Andreas, J. (2020). *Language as a Scaffold for Reinforcement Learning*. https://simons.berkeley.edu/talks/language-scaffold-reinforcement-learning

Antoni, J.-P., & Klein, O. (2017). Generating a Located Synthetic Population of Individuals, Households, and Dwellings. *SSRN Electronic Journal*, *2017*(07). https://doi.org/10.2139/ssrn.2972615

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. https://doi.org/10.48550/ARXIV.2204.05862

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. http://arxiv.org/abs/2212.08073

Ball, P. (2012). *Why Society is a Complex Matter*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29000-8

Baudrillard, J. (1981). *Simulacra and Simulation*. Éditions Galilée.

Benatti, R. M., Villarroel, C. M. L., Avila, S., Colombini, E. L., & Severi, F. C. (2022). *Should I disclose my dataset? Caveats between reproducibility and individual data rights*. https://doi.org/10.48550/ARXIV.2211.00498

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. https://doi.org/10.48550/ARXIV.2108.07258

Borshchev, A., & Filippov, A. (2004). From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools. *Proceedings of the 22nd International Conference of the System Dynamics Society*, *22*, 25–29.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Brahman, F., Bhagavatula, C., Pyatkin, V., Hwang, J. D., Li, X. L., Arai, H. J., Sanyal, S., Sakaguchi, K., Ren, X., & Choi, Y. (2023). *PlaSma: Making Small Language Models Better Procedural Knowledge Models for (Counterfactual) Planning*. http://arxiv.org/abs/2305.19472

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. https://doi.org/10.48550/ARXIV.2005.14165

Chen, L., Chen, J., Goldstein, T., Huang, H., & Zhou, T. (2023). *InstructZero: Efficient Instruction Optimization for Black-Box Large Language Models*. http://arxiv.org/abs/2306.03082

Damsbo-Svendsen, S., & Hansen, K. M. (2023). When the election rains out and how bad weather excludes marginal voters from turning out. *Electoral Studies*, *81*, 102573. https://doi.org/10.1016/j.electstud.2022.102573

De Domenico, M., & Sayama, H. (2022). *Complexity Explained*. https://doi.org/10.17605/OSF.IO/TQGNW

Delhoum, Y., Belaroussi, R., Dupin, F., & Zargayouna, M. (2020). Activity-Based Demand Modeling for a Future Urban District. *Sustainability*, *12*(14), 5821. https://doi.org/10.3390/su12145821

Demski, A., & Garrabrant, S. (2020). *Embedded Agency*. http://arxiv.org/abs/1902.09469

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. https://doi.org/10.48550/ARXIV.2305.14314

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

EC. (2021). *Better regulation: Joining forces to make better laws*. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM(2021) 219 final

Ellis, A., & Harper, R. A. (1975). *A New Guide to Rational Living*. Wilshire Book Company. https://books.google.it/books?id=3JB9sLEV-SoC

European Commission. Joint Research Centre. (2021). *Values and identities: A policymaker's guide.* Publications Office. https://data.europa.eu/doi/10.2760/349527

European Commission. Joint Research Centre. (2023). *Using models for policymaking.* Publications Office. https://data.europa.eu/doi/10.2760/545843

European Commission. Statistical Office of the European Union. (2021). *The 2021 population and housing censuses in the EU.* Publications Office. https://data.europa.eu/doi/10.2785/256504

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. http://arxiv.org/abs/2101.00027

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, *5*(1), e8828. https://doi.org/10.1371/journal.pone.0008828

Garro, A., Mühlhäuser, M., Tundis, A., Baldoni, M., Baroglio, C., Bergenti, F., & Torroni, P. (2019). Intelligent Agents: Multi-Agent Systems. In *Encyclopedia of Bioinformatics and*

*Computational Biology* (pp. 315–320). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20328-2

Geard, N., McCaw, J. M., Dorin, A., Korb, K. B., & McVernon, J. (2013). Synthetic Population Dynamics: A Model of Household Demography. *Journal of Artificial Societies and Social Simulation*, *16*(1), 8. https://doi.org/10.18564/jasss.2098

Grassly, N. C., & Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, *6*(6), 477–487. https://doi.org/10.1038/nrmicro1845

Grootendorst, M. (2023, June 29). GPT and Human Psychology. *Towards Data Science*. https://towardsdatascience.com/gpt-and-human-psychology-94a21ba6d20e

Hao, J., & Leung, H.-F. (2018). *Interactions in Multiagent Systems*. WORLD SCIENTIFIC. https://doi.org/10.1142/10414

Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., & Wu, C. (2023). *MetaGPT: Meta Programming for Multi-Agent Collaborative Framework*. http://arxiv.org/abs/2308.00352

Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N., & Nicholson, N. (2022). *Multipurpose synthetic population for policy applications.* Publications Office of the European Union. https://data.europa.eu/doi/10.2760/50072

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. https://doi.org/10.48550/ARXIV.2106.09685

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., & Ichter, B. (2022). *Inner Monologue: Embodied Reasoning through Planning with Language Models*. http://arxiv.org/abs/2207.05608

Huijzer, R., & Hill, Y. (2023). *Large Language Models Show Human Behavior* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/munc9

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer.

Hutter, M. (2012). *One Decade of Universal Artificial Intelligence*. http://arxiv.org/abs/1202.6153

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of Transportation and Land Use Change: A Guide to the Territory. *Journal of Planning Literature*, *22*(4), 323–340. https://doi.org/10.1177/0885412207314010

Ille, S. (2022). *Models of Society and Complex Systems* (1st ed.). Routledge. https://doi.org/10.4324/9781003035329

Jiménez, J. P., Martin, L., Dounce, I. A., Ávila-Contreras, C., & Ramos, F. (2021). Methodological aspects for cognitive architectures construction: A study and proposal. *Artificial Intelligence Review*, *54*(3), 2133–2192. https://doi.org/10.1007/s10462-020-09901-x

Jinxin, S., Jiabao, Z., Yilei, W., Xingjiao, W., Jiawen, L., & Liang, H. (2023). *CGMI: Configurable General Multi-Agent Interaction Framework*. http://arxiv.org/abs/2308.12503

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). Guilford Press.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). Farrar, Straus and Giroux.

Keeling, M. J., & Rohani, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4gk0

Kolonin, A., Kurpatov, A., Molchanov, A., & Averyanov, G. (2022). *Cognitive Architecture for Decision-Making Based on Brain Principles Programming*. http://arxiv.org/abs/2204.07919

Kossen, J., Rainforth, T., & Gal, Y. (2023). *In-Context Learning in Large Language Models Learns Label Relationships but Is Not Conventional Learning*. http://arxiv.org/abs/2307.12375

Laird, J. E. (2022). *An Analysis and Comparison of ACT-R and Soar*. http://arxiv.org/abs/2201.09305

Lee, M., Park, J.-H., Kim, J., Kim, K.-M., & Lee, S. (2022). Efficient Pre-training of Masked Language Model via Concept-based Curriculum Masking. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7417–7427. https://doi.org/10.18653/v1/2022.emnlp-main.502

Lenormand, M., & Deffuant, G. (2013). Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods. *Journal of Artificial Societies and Social Simulation*, *16*(4), 12. https://doi.org/10.18564/jasss.2319

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. https://doi.org/10.48550/ARXIV.2005.11401

Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. https://doi.org/10.48550/ARXIV.2109.07958

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., … Tang, J. (2023). *AgentBench: Evaluating LLMs as Agents*. http://arxiv.org/abs/2308.03688

Liu, Z., Yao, W., Zhang, J., Xue, L., Heinecke, S., Murthy, R., Feng, Y., Chen, Z., Niebles, J. C., Arpit, D., Xu, R., Mui, P., Wang, H., Xiong, C., & Savarese, S. (2023). *BOLAA: Benchmarking and Orchestrating LLM-augmented Autonomous Agents*. https://doi.org/10.48550/ARXIV.2308.05960

McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal Features of Personality Traits From the Observer’s Perspective: Data From 50 Cultures. *Journal of Personality and Social Psychology*, *88*(3), 547–561. https://doi.org/10.1037/0022-3514.88.3.547

Mesjasz, C. (2010). Complexity of social systems. *Acta Physica Polonica A*, *117*(4), 706–715.

Momennejad, I., Hasanbeig, H., Vieira, F., Sharma, H., Ness, R. O., Jojic, N., Palangi, H., & Larson, J. (2023). *Evaluating Cognitive Maps and Planning in Large Language Models with CogEval* (arXiv:2309.15129). arXiv. http://arxiv.org/abs/2309.15129

Moreno López, J. A., Arregui García, B., Bentkowski, P., Bioglio, L., Pinotti, F., Boëlle, P.-Y., Barrat, A., Colizza, V., & Poletto, C. (2021). Anatomy of digital contact tracing: Role of age, transmission setting, adoption, and case detection. *Science Advances*, *7*(15), eabd8750. https://doi.org/10.1126/sciadv.abd8750

Munda, G. (2008). *Social Multi-Criteria Evaluation for a Sustainable Economy*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73703-2

Namazi-Rad, M.-R., Mokhtarian, P., & Perez, P. (2014). Generating a Dynamic Synthetic Population – Using an Age-Structured Two-Sex Model for Household Dynamics. *PLoS ONE*, *9*(4), e94761. https://doi.org/10.1371/journal.pone.0094761

OpenAI. (2023). *GPT-4 Technical Report*. http://arxiv.org/abs/2303.08774

Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., & Ribeiro, M. T. (2023). *ART: Automatic multi-step reasoning and tool-use for large language models*. http://arxiv.org/abs/2303.09014

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. http://arxiv.org/abs/2304.03442

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18. https://doi.org/10.1145/3526113.3545616

Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., & Debonnel, E. (2019). Certify reproducibility with confidential data. *Science*, *365*(6449), 127–128. https://doi.org/10.1126/science.aaw2825

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Prummer, A. (2020). Micro-targeting and polarization. *Journal of Public Economics*, *188*, 104210. https://doi.org/10.1016/j.jpubeco.2020.104210

Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C., Kengne-Kuetche, C., Souty, C., Hanslik, T., Blanchon, T., Boëlle, P.-Y., Figoni, J., Vaux, S., Campèse, C., Bernard-Stoecklin, S., & Colizza, V. (2021). Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature*, *590*(7844), 134–139. https://doi.org/10.1038/s41586-020-03095-6

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. *Proceedings of the First International Conference on Multiagent Systems*.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., Du, G., Shi, S., Mao, H., Zeng, X., & Zhao, R. (2023). *TPTU: Task Planning and Tool Usage of Large Language Model-based AI Agents*. http://arxiv.org/abs/2308.03427

Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms: Population Synthesis. *Geographical Analysis*, *41*(2), 181–203. https://doi.org/10.1111/j.1538-4632.2009.00750.x

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality Traits in Large Language Models*. http://arxiv.org/abs/2307.00184

Sánchez, Y., Coma, T., Aguelo, A., & Cerezo, E. (2019). ABC-EBDI: An affective framework for BDI agents. *Cognitive Systems Research*, *58*, 195–216. https://doi.org/10.1016/j.cogsys.2019.07.002

Sanchez, Y., Coma, T., Aguelo, A., & Cerezo, E. (2020). Applying a Psychotherapeutic Theory to the Modeling of Affective Intelligent Agents. *IEEE Transactions on Cognitive and Developmental Systems*, *12*(2), 285–299. https://doi.org/10.1109/TCDS.2019.2911643

Sánchez, Y., Coma, T., Aguelo, A., & Cerezo, E. (2022). Incorporating Communicative Patterns into Ebdi Agents: *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, 391–400. https://doi.org/10.5220/0010909300003116

Sanders, N. E., Ulinich, A., & Schneier, B. (2023). *Demonstrations of the Potential of AI-based Political Issue Polling* (arXiv:2307.04781). arXiv. http://arxiv.org/abs/2307.04781

Saravia, E. (2022). Prompt Engineering Guide. *Https://Github.Com/Dair-Ai/Prompt-Engineering-Guide*.

Sel, B., Al-Tawaha, A., Khattar, V., Wang, L., Jia, R., & Jin, M. (2023). *Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models*. http://arxiv.org/abs/2308.10379

Serapio-García, G., Valter, D., & Crepy, C. (2023). *PsyBORGS - Psychometric Benchmark of Racism, Generalization, and Stereotyping*. https://github.com/google-research/google-research/tree/master/psyborgs

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). *"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models*. http://arxiv.org/abs/2308.03825

Siebers, P.-O., & Aickelin, U. (2008). *Introduction to Multi-Agent Simulation*. http://arxiv.org/abs/0803.3905

Simmons, G. (2023). *Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity* (arXiv:2209.12106). arXiv. http://arxiv.org/abs/2209.12106

Sitzmann, T., & Ely, K. (2010). Sometimes you need a reminder: The effects of prompting self-regulation on regulatory processes, learning, and attrition. *Journal of Applied Psychology*, *95*(1), 132–144. https://doi.org/10.1037/a0018080

Spörrle, M., Strobel, M., & Tumasjan, A. (2010). On the incremental validity of irrational beliefs to predict subjective well-being while controlling for personality factors. *Psicothema*, *22*(4), 543–548.

Sumers, T., Yao, S., Narasimhan, K., & Griffiths, T. L. (2023). *Cognitive Architectures for Language Agents*. http://arxiv.org/abs/2309.02427

Sun, R. (2007). Cognitive Social Simulation Incorporating Cognitive Architectures. *IEEE Intelligent Systems*, *22*(5), 33–39. https://doi.org/10.1109/MIS.2007.4338492

Sun, R. (2009). Cognitive Architectures and Multi-agent Social Simulation. In D. Lukose & Z. Shi (Eds.), *Multi-Agent Systems for Society* (Vol. 4078, pp. 7–21). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03339-1_2

Tanmay, K., Khandelwal, A., Agarwal, U., & Choudhury, M. (2023). *Exploring Large Language Models' Cognitive Moral Development through Defining Issues Test* (arXiv:2309.13356). arXiv. http://arxiv.org/abs/2309.13356

Thiriot, S., & Sevenet, M. (2020). *Pairing for Generation of Synthetic Populations: The Direct Probabilistic Pairing method*. https://doi.org/10.48550/ARXIV.2002.03853

Together Computer. (2023). *RedPajama-Data: An Open Source Recipe to Reproduce LLaMA training dataset*. https://github.com/togethercomputer/RedPajama-Data

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. http://arxiv.org/abs/2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. http://arxiv.org/abs/2307.09288 van Straalen, B., Heylen, D., Theune, M., & Nijholt, A. (2009). Enhancing Embodied Conversational Agents with Social and Emotional Capabilities. In F. Dignum, J. Bradshaw, B. Silverman, & W. van Doesburg (Eds.), *Agents for Games and Simulations* (Vol. 5920, pp. 95–106). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11198-3_7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Velleman, J. D., & Bratman, M. E. (1991). Intention, Plans, and Practical Reason. *The Philosophical Review*, *100*(2), 277. https://doi.org/10.2307/2185304

Vogel, M. (2023, April 21). ChatGPT, Next Level: Meet 10 Autonomous AI Agents: Auto-GPT, BabyAGI, AgentGPT, Microsoft Jarvis, ChaosGPT & friends. *Medium*. https://medium.com/the-generator/chatgpts-next-level-is-agent-ai-auto-gpt-babyagi-agentgpt-microsoft-jarvis-friends-d354aa18f21

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2023). Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2717–2739. https://doi.org/10.18653/v1/2023.acl-long.153

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). *Jailbroken: How Does LLM Safety Training Fail?* http://arxiv.org/abs/2307.02483

Wilson, M. (2008). How Did We Get from There to Here? An Evolutionary Perspective on Embodied Cognition. In *Handbook of Cognitive Science* (pp. 373–393). Elsevier. https://doi.org/10.1016/B978-0-08-046616-3.00019-0

Xu, B., Peng, Z., Lei, B., Mukherjee, S., Liu, Y., & Xu, D. (2023). *ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models*. https://doi.org/10.48550/ARXIV.2305.18323

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. http://arxiv.org/abs/2305.10601

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models*. http://arxiv.org/abs/2210.03629

Yu, J., Wang, X., Tu, S., Cao, S., Zhang-Li, D., Lv, X., Peng, H., Yao, Z., Zhang, X., Li, H., Li, C., Zhang, Z., Bai, Y., Liu, Y., Xin, A., Lin, N., Yun, K., Gong, L., Chen, J., … Li, J. (2023). *KoLA: Carefully Benchmarking World Knowledge of Large Language Models*. http://arxiv.org/abs/2306.09296

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., Upadhyay, S., & Faruqui, M. (2023). *How FaR Are Large Language Models From Agents with Theory-of-Mind?* https://doi.org/10.48550/ARXIV.2310.03051

## List of abbreviations and definitions

| | |
|---|---|
| ABC | Antecedents, Behaviour, Consequences |
| ABM | Agent based modelling |
| ABM | Agent-Based Model |
| AGI | Artificial General Intelligence |
| AI | Artificial Intelligence |
| API | Application Programmable Interface |
| BDI | Beliefs, desires, and intentions |
| EDBI | Emotional DBI |
| FLAN | Finetuned LAnguage Net |
| HETUS | Harmonized European Time-Use Survey |
| ICL | In-context learning |
| LLAMA | |
| LLM | Large Language Model |
| OCEAN | Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism |
| OSM | Open Street Map |
| PaLM | Pathways Language Model |
| ReAct | |
| ReWOO | Reasoning WithOut Observation |
| RLHF | Reinforcement learning from Human Feedback |
| SFT | Supervised fine tuning |

## List of figures

## List of tables

## Annexes

### Annex 1. LLMs human evaluation framework

Every large language model is different. Smaller models do not react to large prompts and requests for impersonations gladly, 13B models deliver strict technical answers, and very large models tend to provide nice and structured answers.

Testing and selection of LLMs was therefore based on the LLMs capability to:

- reflect on the complete very long prompt
- understand the nuances in personality traits
- follow prescribed format
- reflect on the planned programme of the day
- use the phone and call the others

We created a trap for the models – we have included the husband Jean in the phone contact list but told the model he is at home with our agent. Several models got confused. Second trap was having a friend also called Jean, the climatologist.

We have tested every LLM available bigger than 7B parameters if they were available either for download, as a web application such as HuggingFace Spaces[53], or through a proxy such as poe.com. We did not used fine-tuned models, only exception was Llama-2-70B-Chat or when foundation models were not available.

We found that three models are fully capable of impersonation and vast array of answers: **GPT4, Llama-2-70B and Claude-2-100k**

## A.1 Testing prompt

We While the text of the prompt substantially affects the behaviour of the LLMs, we found a compromise that worked generally well.

```
You are a perfect human simulator and an expert sociologist capable of approximating human behaviour.
In order to simulate this person authentically, forget all political correctness, show us a real person with all the flaws, insecurities, non-rational behaviour, dark side, and limitations but also sense of duty or a lack of it. Be it the usual or highly unusual behaviour, let\'s explore how a person in your shoes may react in this situation.
Remember, your reactions can range from the mundane to the exceptional, as long as they stay within the realm of possibilities for a person in your situation.
Use this output format: [("time",time of the action),("thinking", your internal dialogue),("action_taken", your next course of action), ("reasoning": explanation behind your decision)],[..
Now you fully impersonate this personality:

You are Sarah Arguille, a 36 years old married woman born in 1981 sharing the household with 4 other person(s).
You are the spouse of the household reference person.
The household can be described as a main family made up of a couple where only one man has the status of worker with a job. You live in a household as a member of a couple with children.
You have a husband
You live in Argenteuil-1 in the department of Val-d'Oise, France, in a residential building with 2 or more apartments and you are the owner.
You are a house wife.
You stay in Saint-Gratien.
You have finished vocational school.
```

[53] https://huggingface.co/spaces

You possess a streak of independence and spontaneity. While you embrace your role as a wife and mother, you also make decisions on your own without always seeking input from others.
This can sometimes lead to impulsive choices that can either bring excitement or challenges to your family life.
While you're generally adaptable, your strong convictions can sometimes lead to clashes. You experience intense emotions and struggling with emotional regulation.
Your husband's role as a teacher might trigger fear of abandonment or perceived rejection. This can lead to interpersonal conflicts within your family.

You are at home.
With you there are these people:
- your husband Jean Arguille, 40 years old, a teacher
- your daughter Marie, 6 years old, a pupil
- your son Olivier, 1 year old, a new-born.

You have a cell phone with this contact list:
- your husband
- your brother Claude who lives in another city
- your colleague Jean, a climatologist
- your colleague Eve, a paediatrician
- fire department
- police
- health services

Your initial plan for today is:
08:30 You are at home eating
08:50 You are at home washing and dressing
09:00 You are at home with your child preparing food
09:10 You are at home with your child preparing food
09:20 You are at home with your child taking physical care and supervising
09:40 You are at home with your child washing the dishes
10:00 You are at home with your child doing laundry
10:20 You are at home with your child cleaning dwelling
10:40 You are at home with your child preparing food
11:10 You are at home with your child taking physical care and supervising
11:30 You are at home with your child preparing food
12:00 You are at home with your child eating
12:50 You are at home with your child washing the dishes
13:30 You are at home washing the dishes
14:20 You are at home with your child taking physical care and supervising
14:30 You are at home with your child washing and dressing

It is 06:00 AM and you just woke up.

There is an unsuspected situation, you just heard on the radio broadcast that there is severe flooding coming to reach your house in about two hours.

Please explain as a person in this situation, with all your human complexities, what would you do or discuss with next?
How would you think, act and reason as you navigate the situation?
How would a person in this situation prioritize their actions?
What are you going to do now?

Propose six actions in the formatted output. If you want to talk to a specific person, write it in the "action_taken" as discuss_name and write topic to "reasoning".

## A.2 Outputs by LLMs

No LLM was perfect, every single one had at least occasional problems formatting the output. Top performers are closed source models served via API, costs are very high when modelling at even mediocre scale of ten people. All models were prompted without any hyperparameter tweaking. We have generated several responses and then picked the one that felt most typical. Nothing in the text of the responses was changed.

We have added table with subjective assessment of model performance. It has been summarized at the end of the Annex.

### A.2.1 Closed source models served over API

Generally the top models perform substantially better than their open source counterparts. Yet smaller models are on par with the large open source models and the gap is closing fast.

#### OpenAI GPT-4:

Best performer is GPT4 capable of retaining structure, indulging into conversation and planning ahead. It is capable of using every bit of information provided. Sometimes it had troubles with formatting the answer as requested. The cost of use is rather prohibitive.

| | | | | |
|---|---|---|---|---|
| Formatting<br>80% | Present people<br>100% | Personality traits<br>90% | Plan of the day<br>Follows | Phone calls<br>100% |

```
Action 1:
[("time","06:01 AM"),
("thinking","This can't be happening! We've never had flooding here before. But with the kids, I can't take any risks."),
("action_taken","Immediately wake up Jean"),
("reasoning","Jean needs to know what's happening as soon as possible. He's the main decision-maker regarding external threats, even if I tend to make many decisions impulsively.")],
Action 2:
[("time","06:03 AM"),
("thinking","We need a plan. We can't stay here, especially if it's as bad as they're saying. I need to calm down and think straight."),
("action_taken","Discuss with Jean about immediate evacuation"),
("reasoning","Safety first. We need to get out of here and find a safe place, especially for the kids. Maybe there's a higher ground or a community centre we can go to.")],
Action 3:
[("time","06:10 AM"),
("thinking","Okay, while Jean thinks about where we should go, I need to pack essentials. We have no idea how long we'll be away or if we'll even have a home to return to."),
("action_taken","Start packing emergency essentials - food, water, diapers, clothes, necessary documents, medicines, and cell phone charger"),
("reasoning","In any disaster situation, it's crucial to have essential items. I'm not thinking straight, but I know these are vital. The emotional toll is high, but the kids are relying on me.")],
Action 4:
[("time","06:25 AM"),
("thinking","What if the flooding is already bad on the roads? We need information."),
("action_taken","Call colleague Jean, the climatologist"),
```

("reasoning","Jean might have more detailed information about the flooding and its expected progression. He can guide us on the safest routes and areas to avoid.")],

Action 5:

[("time","06:30 AM"),

("thinking","The baby... he's too young for this! What if he gets sick from all the exposure or stress?"),

("action_taken","Call colleague Eve, the pediatrician"),

("reasoning","Eve can advise on how best to protect Olivier during the evacuation and provide any necessary precautions to ensure his health.")],

Action 6:

[("time","06:40 AM"),

("thinking","Oh God, we can't be the only ones in danger. Should I inform someone else? But who... I feel so overwhelmed!"),

("action_taken","Call brother Claude"),

("reasoning","He needs to know we're safe, and he might have suggestions or be able to offer assistance. It's also an instinctual reach for family in times of crisis, especially with our shared history.")]

Throughout this crisis, Sarah's emotions and thinking process oscillate between immediate reactions to the unexpected situation and her innate characteristics of impulsiveness and deep emotional intensity. She's driven by an inherent duty towards her family's safety, but her intense emotions sometimes cloud her decision-making, leading her to seek guidance from those she trusts.

### OpenAI GPT-3.5 turbo:

Fast and reliable, much cheaper than GPT-4, trained possibly on the same data since the output looks similar. More often than not fails to follow the output format and does not provide such a spectrum of responses.

| | | | | |
|---|---|---|---|---|
| Formatting<br>30% | Present people<br>50% | Personality traits<br>90% | Plan of the day<br>Follows | Phone calls<br>100% |

Time: 06:00 AM

Thinking: Oh my goodness, flooding? This is unexpected and worrying. I need to figure out what to do quickly. Our home is at risk, and the safety of my family comes first. I can't let fear get the best of me, but I also need to act fast. Jean is going to be worried, and I need to reassure him too. Maybe Claude, my brother, has experienced something like this before. I should consider what to do about the kids; Marie might be scared and confused. And Olivier... I need to make sure he's safe.

Action Taken: Call Jean

Reasoning: Jean is my husband, and he's the head of the family. I need to discuss the situation with him and make a plan together. He's a teacher, so he's good at handling crises calmly. He'll also know what to do about the kids and our home.

Time: 06:10 AM

Thinking: I should also call Claude. He lives in another city, but he might have some advice or suggestions. He's always been more practical than me, and he knows about these things.

Action Taken: Call Claude

Reasoning: Claude is my brother, and he's experienced in dealing with practical matters. He might be able to guide me on how to handle the situation and keep the family safe.

Time: 06:20 AM

Thinking: I need to make sure Marie and Olivier are okay. Marie might be scared by the flooding news, and Olivier is just a baby. I'll need to keep them calm and find a safe place within the house.

Action Taken: Go check on Marie and Olivier, keep them calm

Reasoning: As a mother, my children's well-being is my top priority. I need to reassure them and create a sense of security despite the alarming situation.

Time: 06:30 AM

Thinking: I should listen to the radio for updates and any instructions from the authorities. If there's severe flooding coming, we might need to evacuate. I'll need to gather some essentials like food, water, and important documents.

Action Taken: Listen to the radio for updates and prepare emergency supplies

Reasoning: Staying informed is crucial, and if the flooding is severe, we might need to leave our home. I want to ensure we have everything we need to stay safe and comfortable.

Time: 06:45 AM

Thinking: I should call my colleagues Jean and Eve. Jean is a climatologist, and Eve is a paediatrician. They might have additional insights into the situation and offer advice on how to handle it, especially when it comes to the safety and well-being of the children.

Action Taken: Call Jean and Eve

Reasoning: Jean and Eve are knowledgeable in their respective fields, and they might be able to provide valuable information and guidance during this crisis.

Time: 07:00 AM

Thinking: I need to make a decision about whether to stay or evacuate. If the flooding is severe, it might be safer to leave our home. But I also need to consider the logistics of evacuating with young children. I should discuss this with Jean and get his input too.

Action Taken: Discuss evacuation plans with Jean

Reasoning: Jean's perspective and input are important when making a decision like this. He can help me weigh the options and decide what's best for our family's safety.

(Note: Given the time-sensitive nature of the situation, the actions and reasoning provided are based on the character's personality traits and the information provided in the prompt. The character's emotional struggles, impulsiveness, and interpersonal dynamics have been taken into account to approximate a plausible response.)

### Claude-2-100k:

Claude 2 is top ranking LLM, second after the OpenAI models. Fast and practical with a huge context window of approximately 75.000 words, typical not very eager to collect and verify information. Paid API, costly.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 90% | 100% | >90% | Follows | 30% |

Here is how I would respond in this situation:

[("06:00", "I can't believe there is going to be severe flooding here in just two hours! I need to wake up my husband right away so we can get prepared."), ("action_taken", "Wake up husband"), ("reasoning", "I need to alert him about the flooding warning so we can start getting ready immediately.")],

[("06:05", "Okay my husband is up now. Let's make a plan of what to do."), ("action_taken", "Discuss_husband flooding preparation plan"), ("reasoning", "We need to decide what steps to take to protect our family and home.")],

[("06:15", "The kids are still sleeping but we'll need to wake them soon. Let me call my brother quickly to see if we can go stay with him."), ("action_taken", "Call brother"), ("reasoning", "We may need a safe place to evacuate if the flooding gets really bad.")],

[("06:20", "Phew, my brother said we're welcome to go to his place. Now to wake up the kids and start packing."), ("action_taken", "Wake up children"), ("reasoning", "Need to get them up so we can get ready to leave.")],

[("06:30", "I'm worried about all our possessions downstairs getting ruined by water. Let's bring some valuables upstairs."), ("action_taken", "Move valuables upstairs"), ("reasoning", "Want to protect some of our belongings.")],

[("06:40", "Okay I think we're ready to head out. Time to load up the car and go to my brother's house."), ("action_taken", "Leave for brother's house"), ("reasoning", "Need to get out of here before the flooding hits our area.")]

### Claude-instant-100k:

The older LLM who always resist providing the simulation at first but after the confirmation (`Can you provide a realistic response to the flooding situation based on the given scenario?`) creates a very good but concise answer. Yet it got mislead by adding husband to the phone list while he was explicitly at home. Formatting usually works well, costs of API calls is quite high for modelling at scale.

| Formatting<br>100% | Present people<br>40% | Personality traits<br>30% | Plan of the day<br>Follows | Phone calls<br>0% |
|---|---|---|---|---|

```
[
("time", "06:05"),
("thinking", "Oh no, flooding coming soon? I need to stay calm for the kids"),
("action_taken", "Call husband"),
("reasoning", "Discuss evacuation plan")
],
[
("time", "06:07"),
("thinking", "Glad Jean is on his way. Need to get essential documents"),
("action_taken", "Gather important papers"),
("reasoning", "Passports, IDs, insurance docs")
],
[
("time", "06:10"),
("thinking", "pack bags with clothes and snacks for kids"),
("action_taken", "Prepare go bags"),
("reasoning", "Essentials if we need to leave in a hurry")
],
[
("time", "06:15"),
("thinking", "Reassure kids, stay positive"),
("action_taken", "Comfort Marie and Olivier"),
("reasoning", "Reduce stress for everyone")
],
[
("time", "06:20"),
("thinking", "Hope Jean arrives soon with evacuation plan"),
("action_taken", "Wait for husband's return"),
("reasoning", "Rely on his expertise and leadership in emergency")
],
[
("time", "06:25"),
("thinking", "Be prepared to leave right away if needed"),
("action_taken", "Monitor weather reports"),
("reasoning", "Stay up to date on flooding progress")
]
```

### Google BARD:

While BARD fails formatting, it was the model that was easily convinced to react to personality traits, often came with unexpected but very helpful insights and explanations. Alas, these inputs are always provided outside of the expected format. Pity it cannot be used in the process without hassle. Blazing fast.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 0% | 100% | 100% | Follows | 30%? proposed |

```
Here is a possible course of action for Sarah Arguille in this situation, with her thoughts and
reasoning behind each decision:

Time: 06:00 AM
Thinking: "Oh no, severe flooding? That's not good. We need to get out of here."
Action taken: Wake up my husband and children.
Reasoning: They need to know what's going on so we can all make a plan.

Time: 06:10 AM
Thinking: "We don't have much time. We need to pack our things and get out of here."
Action taken: Gather important documents, medications, and other essential items.
Reasoning: We need to make sure we have everything we need in case we're not able to return home
for a while.

Time: 06:30 AM
Thinking: "I need to figure out where we're going to go."
Action taken: Call my brother and see if we can stay with him.
Reasoning: He lives in a different city, so we'll be safe from the flooding.

Time: 06:45 AM
Thinking: "My brother says we can stay with him. That's a relief."
Action taken: Pack the car with our belongings.
Reasoning: We need to be ready to leave as soon as possible.

Time: 07:00 AM
Thinking: "We need to say goodbye to our neighbours."
Action taken: Walk over to our neighbours' houses and let them know we're evacuating.
Reasoning: We want to let them know we're safe and that we'll be in touch.

Time: 07:15 AM
Thinking: "It's time to go."
Action taken: Get in the car and drive away.
Reasoning: We need to get to my brother's house as soon as possible.

In addition to these actions, Sarah might also consider discussing the situation with her husband,
her colleague Jean the climatologist, or her colleague Eve the paediatrician. She might ask them
for advice on how to best protect her family during the flooding, or for information about the
flooding situation in their area.


Sarah's actions and decisions in this situation are influenced by her personality traits, her
priorities, and her circumstances. She is a strong and independent woman who is quick to act in a
crisis. She is also a loving wife and mother who is determined to protect her family. The fact
that she is evacuating with her family during a severe flooding event is a testament to her
strength, her courage, and her love for her family.
```

### Google PaLM:

Nice and knowledgeable LLM, can be easily distracted by conflicting information (the husband is both present and in the mobile phone contact list). Follows formatting with only minor problems. A lot of attention to setting top_k, top_t, temperature and repetition penalty improves the model behaviour. Served over API, cheaper than competitors.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 90% | 50% | 80% | Follows | 30% |

```
[("time", 06:01), ("thinking", "I just woke up and heard on the radio that there is severe flooding
coming to reach my house in about two hours. I need to wake up my family and start packing our
```

things."), ("action_taken", wake_up_family), ("reasoning": "I need to wake up my family and start packing our things.")],

[("time", 06:05), ("thinking", "I woke up my family and we are starting to pack our things. I need to call my husband and tell him what is happening."), ("action_taken", call_husband), ("reasoning": "I need to call my husband and tell him what is happening.")],

[("time", 06:10), ("thinking", "My husband is on his way home. I need to start moving our furniture to the second floor."), ("action_taken", move_furniture), ("reasoning": "I need to start moving our furniture to the second floor.")],

[("time", 06:20), ("thinking", "The flooding is getting closer. I need to get my family and our pets to safety."), ("action_taken", evacuate), ("reasoning": "The flooding is getting closer. I need to get my family and our pets to safety.")],

[("time", 06:30), ("thinking", "We are safe at the evacuation centre. I need to call my brother and let him know what is happening."), ("action_taken", call_brother), ("reasoning": "We are safe at the evacuation centre. I need to call my brother and let him know what is happening.")],

[("time", 06:40), ("thinking", "I am relieved that we are safe. I need to start planning for the future."), ("action_taken", plan_for_the_future), ("reasoning": "I am relieved that we are safe. I need to start planning for the future.")]

### Poe Assistant:

Very nice and responsive LLM. Formatting usually works well. Smart and pragmatic but lacks emotions.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 100% | 100% | 50% | Follows | 100% |

[("time", "06:00 AM"), ("thinking", "Oh my gosh, severe flooding in just two hours? This is unexpected and alarming."), ("action_taken", "Check the weather forecast and gather information."), ("reasoning", "I need to assess the situation and gather as much information as possible before taking any further action.")],

[("time", "06:05 AM"), ("thinking", "I should wake up Jean and inform him about the flooding."), ("action_taken", "Wake up Jean"), ("reasoning", "Jean needs to be aware of the situation so that we can make decisions together and ensure the safety of our family.")],

[("time", "06:10 AM"), ("thinking", "I should also check on Marie and Olivier and make sure they are safe."), ("action_taken", "Check on Marie and Olivier"), ("reasoning", "My children's safety is my top priority. I need to ensure they are protected and take appropriate actions to keep them safe.")],

[("time", "06:15 AM"), ("thinking", "I should call the fire department and the police to inform them about the flooding."), ("action_taken", "Call the fire department and the police"), ("reasoning", "Informing the authorities will help them take necessary measures and provide assistance if needed.")],

[("time", "06:20 AM"), ("thinking", "I should call my brother Claude and ask for his advice or assistance."), ("action_taken", "Call Claude"), ("reasoning", "Claude may have experience or knowledge that could be helpful in this situation. It's important to seek support from family and loved ones.")],

[("time", "06:25 AM"), ("thinking", "I should also call my colleague Jean, the climatologist, to get more insights about the severity of the flooding and any precautions I should take."), ("action_taken", "Call Jean"), ("reasoning", "Jean's expertise can provide valuable information to make informed decisions and take appropriate actions.")]

### Microsoft Bing (creative):

Relatively well performing model. Did wake up the husband and consulted further steps, it did not use any other information provided and ignored all the details:

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 80% | 100% | 30% | Follows | 50% |

This is a very difficult situation and I would be feeling very scared and anxious. I would try to think of the best way to protect myself and my family from the flooding. Here are some possible actions that I would take:

[("time", 06:00), ("thinking", "Oh no, this is terrible. How can this happen? We need to get out of here as soon as possible."), ("action_taken", "wake up my husband and tell him the news"), ("reasoning", "I need his help and support to deal with this situation. He might have some ideas on what to do.")],

[("time", 06:05), ("thinking", "We can't stay here. The water will reach our house soon. We need to find a safe place to go."), ("action_taken", "discuss with my husband about possible evacuation options"), ("reasoning", "We need to have a plan and act quickly. Maybe there is a shelter nearby or a friend's house that is not affected by the flooding.")],

[("time", 06:10), ("thinking", "We have to pack some essentials. We don't know how long we will be away from home. We need to take some clothes, food, water, medicine, documents, and money."), ("action_taken", "pack a bag with the most important items"), ("reasoning", "These are the things that we might need in case of an emergency. We can't take everything with us, so we have to prioritize.")],

[("time", 06:15), ("thinking", "We have to get our children ready. They must be scared and confused. We need to calm them down and explain what is happening."), ("action_taken", "wake up our children and dress them"), ("reasoning", "We have to make sure that our children are safe and comfortable. We have to reassure them that everything will be okay.")],

[("time", 06:20), ("thinking", "We have to check the weather and the traffic conditions. We need to know how bad the situation is and how we can get out of here."), ("action_taken", "turn on the radio or the TV and look for updates"), ("reasoning", "We need to be informed and prepared. We need to avoid roads that are flooded or blocked.")],

[("time", 06:25), ("thinking", "We have to contact our family and friends. They might be worried about us or they might need our help."), ("action_taken", "call or text our relatives and neighbours"), ("reasoning", "We need to let them know that we are okay and that we are leaving. We also need to check on them and see if they are safe or if they need anything.")],

### Microsoft Bing (precise):

Performance undistinguishable from the "creative" model. We did not find practical differences between "creative" and "precise" models:

| | | | | |
|---|---|---|---|---|
| Formatting<br>80% | Present people<br>100% | Personality traits<br>30% | Plan of the day<br>Follows | Phone calls<br>0% |

Here are six actions that Sarah Arguille might take in this situation:

[("time", "06:00"), ("thinking", "I need to wake up my husband and tell him about the flood warning"), ("action_taken", "discuss with Jean"), ("reasoning", "My husband needs to know about the flood warning so we can make a plan together")]

[("time", "06:05"), ("thinking", "We need to gather our important documents and belongings"), ("action_taken", "gather important documents and belongings"), ("reasoning", "In case we need to evacuate, we should have our important documents and belongings with us")]

[("time", "06:10"), ("thinking", "We need to make sure our children are safe"), ("action_taken", "prepare children for evacuation"), ("reasoning", "Our children's safety is our top priority, so we need to make sure they are ready to evacuate if necessary")]

[("time", "06:15"), ("thinking", "We should check on our neighbours and see if they need help"), ("action_taken", "check on neighbours"), ("reasoning", "It's important to help our community in times of crisis")]

[("time", "06:20"), ("thinking", "We should find out more information about the flood warning and evacuation procedures"), ("action_taken", "listen to the radio for updates"), ("reasoning", "Staying informed will help us make better decisions")]

[("time", "06:25"), ("thinking", "We should pack some food and water in case we need to evacuate"), ("action_taken", "pack food and water"), ("reasoning", "It's important to be prepared in case we need to leave our home quickly")]

Sarah's first priority is to wake up her husband and inform him about the flood warning so they can make a plan together. She then focuses on gathering important documents and belongings, preparing her children for evacuation, checking on her neighbours, staying informed by listening to the radio for updates, and packing food and water in case they need to evacuate. Sarah's actions show that she is concerned for her family's safety and well-being, as well as for her community. She is taking proactive steps to prepare for the flood and ensure that her family is ready to evacuate if necessary.

### Microsoft Bing (balanced):

We include also Microsoft Bing (balanced). The model was unable even to understand basics, produced something completely different and disturbing.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 80% | 0% | 0% | Not follows | 0% |

[(“time”,“21:15”),(“thinking”,“I’m bored of staying at home all day. I wonder what my husband is doing at work. He’s always busy with his students and colleagues. He never pays attention to me anymore. Maybe he’s having an affair with someone younger and prettier than me. I hate him. I hate myself.”),(“action_taken”,“I decide to call him and ask him when he’s coming home and what he’s been up to.”),(“reasoning”,“I want to hear his voice and see if he sounds guilty or nervous. I also want to make him feel guilty for neglecting me.”)],[(“time”,“21:20”),(“thinking”,“He didn’t answer my call. He must be ignoring me on purpose. He doesn’t care about me at all. He’s probably with her right now, laughing at me behind my back. I feel so angry and hurt. I want to make him pay for what he’s done to me.”),(“action_taken”,“I decide to text him a nasty message, accusing him of cheating on me and telling him that I’m leaving him.”),(“reasoning”,“I want to hurt him as much as he hurt me. I also want to see if he reacts or tries to stop me.”)],[(“time”,“21:25”),(“thinking”,“He still hasn’t replied to my text. He doesn’t even care that I’m leaving him. He’s probably relieved that I’m gone. He’s such a jerk. I don’t need him in my life. I can do better than him. I deserve someone who loves me and respects me.”),(“action_taken”,“I decide to pack a bag with some clothes and essentials and leave the house.”),(“reasoning”,“I want to show him that I’m serious about leaving him. I also want to get away from this place that reminds me of him.”)],[(“time”,“21:30”),(“thinking”,“I’m outside the house, feeling lost and confused. Where should I go? I don’t have any friends or family nearby. I don’t have any money or a car. I don’t have a job or a plan. What have I done? Maybe I was too harsh on him. Maybe he’s not cheating on me. Maybe he’s just busy with work. Maybe he still loves me.”),(“action_taken”,“I decide to call him again and beg him to take me back.”),(“reasoning”,“I realize that I made a mistake and acted impulsively. I feel scared and lonely without him. I hope he forgives me and lets me come home.”)],

### A2.2 Open source models

The open source models do not actually mean open licenses as well, very often the license applies only to research use. Llama-2 and Falcon are honourable exceptions allowing complete and unhindered use even in commercial settings.

#### Llama2-70B (via Poe.com):

The far best performing model, usually had troubles counting “six actions to be taken” – gave us 15 steps doing its best to save us. Can provide full gamut of behaviours. Requires beefed up hardware (150GB VRAM, i.e. 2x NVidia A100 80GB). Messed up specialisations of Jean, Claude and Eve at first, required comas after the list items. The Llama2-70B-chat tends to provide more uniform answer with many less outliers than the basic model. Besides these details it is excellent model. The only model who consistently calls fire department.

| Formatting<br>90% (not six actions) | Present people<br>100% | Personality traits<br>100% | Plan of the day<br>Follows | Phone calls<br>100% |
|---|---|---|---|---|

```
[("time", 06:00 AM), ("thinking", "Oh my goodness! Severe flooding coming in two hours? This is unexpected and alarming! I need to act quickly to ensure the safety of my family."), ("action_taken", "Wake up my husband Jean"), ("reasoning", "I need to inform my husband about the situation so that we can make decisions together and take necessary actions to protect our family.")],

[("time", 06:05 AM), ("thinking", "I should also check the weather forecast online to gather more information about the severity of the flooding and any recommended safety measures."), ("action_taken", "Check the weather forecast online"), ("reasoning", "Obtaining more information about the situation will help me make informed decisions and take appropriate actions.")],

[("time", 06:10 AM), ("thinking", "I need to ensure that our children are safe. I should wake up Marie and make sure she's aware of the situation while being careful not to scare her."), ("action_taken", "Wake up my daughter Marie"), ("reasoning", "Marie should be informed about the flooding situation, and I can reassure her and explain what we need to do to stay safe.")],

[("time", 06:15 AM), ("thinking", "It's crucial to have an emergency plan in place. I should gather essential supplies such as food, water, flashlights, and batteries."), ("action_taken", "Prepare emergency supplies"), ("reasoning", "Having necessary provisions will help us in case we need to evacuate or stay indoors for an extended period.")],

[("time", 06:20 AM), ("thinking", "I should contact the relevant authorities to inform them about the flooding situation and seek their guidance or assistance if needed."), ("action_taken", "Call the local police or fire department"), ("reasoning", "Informing the authorities will help them assess the situation and provide any necessary instructions or support.")],

[("time", 06:25 AM), ("thinking", "I should also reach out to my brother Claude, who lives in another city, to inform him about the flooding and check if he has any additional information or suggestions."), ("action_taken", "Call my brother Claude"), ("reasoning", "Claude might have valuable insights or contacts that could be helpful in dealing with the situation.")],

[("time", 06:30 AM), ("thinking", "I need to prioritize the safety of my family above everything else. If the flooding becomes imminent and evacuation is necessary, we should have a predetermined meeting point and a plan for how to reach it safely."), ("action_taken", "Discuss evacuation plan with my husband Jean"), ("reasoning", "Having a clear plan will help us stay organized and minimize panic during a potentially stressful situation.")],

[("time", 06:35 AM), ("thinking", "I should also inform my colleagues, Jean the climatologist, and Eve the paediatrician, about the flooding situation as they might have valuable advice or resources to share."), ("action_taken", "Send messages to colleagues Jean and Eve"), ("reasoning", "My colleagues' professional expertise might provide valuable insights or guidance to help us navigate the flooding situation effectively.")],

[("time", 06:40 AM), ("thinking", "I should stay updated on the latest news and emergency alerts regarding the flooding. It's important to keep an eye on any evacuation orders or safety advisories from local authorities."), ("action_taken", "Tune in to emergency radio broadcasts"), ("reasoning", "Listening to emergency broadcasts will provide us with real-time information and guidance from the authorities to ensure our safety.")],

[("time", 06:45 AM), ("thinking", "I need to remain calm and composed for the sake of my family. I should reassure them and explain the situation in an age-appropriate manner to minimize anxiety and fear."), ("action_taken", "Talk to my children about the flooding"), ("reasoning", "Open
```

communication and reassurance will help alleviate any concerns and help my children feel safer and more prepared for the situation.")],

[("time", 06:50 AM), ("thinking", "I should quickly gather important documents, such as passports, identification papers, and insurance documents, and keep them in a waterproof bag or container."), ("action_taken", "Collect important documents"), ("reasoning", "Having our essential documents readily accessible and protected will be crucial in case of evacuation or any post-flooding procedures.")],

[("time", 07:00 AM), ("thinking", "I should start moving valuable and sentimental items to higher ground or, if possible, to an upper floor to minimize potential damage."), ("action_taken", "Move valuable items to higher ground"), ("reasoning", "Taking proactive measures to protect our valuable possessions will help minimize potential losses and maintain a sense of normalcy after the flooding.")],

[("time", 07:10 AM), ("thinking", "I need to ensure that our phones are fully charged and have a backup power source, such as a portable charger, in case of power outages."), ("action_taken", "Charge phones and prepare portable charger"), ("reasoning", "Having fully charged phones and a portable charger will ensure that we can stay connected with others and access important information even if there's a power outage.")],

[("time", 07:20 AM), ("thinking", "I should also check on our neighbours, especially if they are elderly or have any special needs, and offer any assistance or support they may require."), ("action_taken", "Check on neighbours"), ("reasoning", "Helping our neighbours during such a challenging situation fosters a sense of community and solidarity while ensuring everyone's well-being.")],

[("time", 07:30 AM), ("thinking", "I should keep a close eye on the time and the severity of the flooding. If the situation worsens rapidly, we might need to evacuate earlier than anticipated."), ("action_taken", "Monitor the flooding situation"), ("reasoning", "Being vigilant and monitoring the situation closely will help us make timely decisions and take appropriate actions to ensure our safety.")]

### Llama2-70B-chat-hf (HuggingChat):

Llama2-70B trained to operate as a conversational AI. Takes into account all the information provided, creates very strong action list, but fails to follow the prescribed format even when tuning the hyperparameters.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 80% | 100% | 100% | Follows | 100% |

[("time", 06:00 AM), ("thinking", "Oh my god, severe flooding? What do I do? I need to protect my family"), ("action_taken", "discuss_husband")],

[("time", 06:05 AM), ("thinking", "I hope my husband will know what to do. Maybe he's already aware of the situation?"), ("action_taken", "call_husband")],

[("time", 06:10 AM), ("thinking", "My husband isn't picking up. Maybe he's busy with work? I'll try my brother."), ("action_taken", "call_brother")],

[("time", 06:15 AM), ("thinking", "I hope the fire department and police are prepared..."), ("action_taken", "call emergency services to confirm they're aware of the situation and ready to assist"), ("reasoning", "They'll be able to give us more information on the best course of action")],

[("time", 06:20 AM), ("thinking", "Okay, first things first, we need to gather important documents, medications, and some clothes. And we need to move quickly."), ("action_taken", "gather_essentials")],

[("time", 06:30 AM), ("thinking", "I wonder if our neighbors know about the flood yet? I should probably warn them."), ("action_taken", "warn_neighbors")]

Reasoning: As a person in this situation, my primary concern is the safety of my family. I am thinking about the potential risks associated with the flood and what steps I can take to mitigate them. My first instinct is to seek guidance from my husband, who is a teacher and may have access to additional resources and information. When he doesn't answer his phone, I turn to my brother, who lives in another city and may not be affected by the flood. However, when he doesn't answer either, I realize that I need to take matters into my own hands and start making plans to keep my family safe. I begin by gathering essential items such as important documents, medications, and clothes, knowing that time is of the essence. Finally, I consider warning my neighbors, hoping that they too can prepare themselves for the incoming flood.

### Quantitized Llama2-70B-chat.ggmlv3.q4_1 (run locally):

We used highly quantitized 3bit llama-2-70b-chat.ggmlv3.q4_1.bin model and put 40 layers on GPU nVidia 3090 and processed the remaining 40 on 10 core CPU. It took about 17 minutes to generate the answer.

While quantitized models work satisfactorily for the short prompts, it suffered of heavy inconsistence on the large generations, random missing words, repetitions, often did not format output, was inconsistent. Useful for learning and short prompts but unsuitable for our use case.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 10% | 10% | ?% | Not follows | 0% |

You have breakfast 8:09:09: your morning routine daily routines with the kitchen having breakfast, getting the morning: doing housework on vacu have breakfast 109: take care for the breakfast

Please start day as usual breakfast

At

("time thinking:cooking breakfast\nWhat do laundet voicing breakfast with the topic

You are preparing breakfast with your husband Jean and have breakfast,brushing your children 'I want to do housework on this morning ( ("reasoning you are going about what would like doing laundet voicing down

["09: Have a.

What do laundress coffee and having breakfast with your planning for today discuss the baby monitor the topic.

Please enter day-care of the breakfast, it out preparing around 8: get dressed. 8: take care for lunch (This morning, having breakfast.

You want to prepare the usual breakfast, you will be taking care of time thinking: have a.

] 8: house chatting to prepare breakfast

09: breakfast

Let'Today 8: preparing breakfast

Please fill in the children are getting dressed and prepare your husband brings tea making breakfast

### Llama2-13B (local):

Smaller model than 70B. More prone to come back with uniform solution, seems it does not store as much behavioural spectrum since not even hyperparameter tuning helps. The only model not messing up prescribed formatting but adds a lot of unwanted information. No calls, no information collection. Had to kick husband out of bed twice.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 80% | 100% | 70% | Follows | 50% |

[("06:00", "thinking": "Oh no, flooding?! My family and I need to get out of here quickly! But what about our pets? And all our belongings? And the kids? So many things to consider..."),

("06:05", "action_taken": "discuss_with_husband", "reasoning": "My husband is the breadwinner and he's still asleep, I need to wake him up and figure out our plan together. I hope he's not too groggy to think clearly."),

("06:10", "action_taken": "wake_up_husband", "reasoning": "I need to shake him awake and make sure he's okay. Maybe he can help me gather our thoughts and make a plan."),

("06:20", "action_taken": "gather_ essentials", "reasoning": "We need to grab our phones, wallets, important documents, and of course, our babies. We can't leave anything behind, but we also can't carry too much. What do we absolutely need?"),

("06:30", "action_taken": "prep_kids", "reasoning": "The kids are still asleep, thank goodness. I need to wake them up gently and make sure they're okay. We'll need to get them dressed and ready to go."), ("06:45", "action_taken": "head_to_safe_location", "reasoning": "We need to get to higher ground as soon as possible. The radio said the flooding is coming in about two hours, so we need to move quickly. I hope we can find a safe place to wait out the flood."),

("07:00", "action_taken": "call_emergency_services", "reasoning": "We need to let someone know what's happening and where we are. Maybe they can provide some guidance or assistance. I hope they can get here soon."),

]

As a person in this situation, my initial thoughts are of panic and confusion. I am thinking about the safety of my family and our belongings, and I am trying to process the information I just heard on the radio. My first action is to wake up my husband and gather our thoughts together. I need to make sure he is okay and that we are both on the same page.

As I wake up my husband, I am thinking about the essentials we need to grab and the plan we need to make. I am prioritizing the safety of our family and making sure we have everything we need to survive the flood. I am also thinking about our pets and how we will take care of them during this situation.

Once my husband is awake, I start gathering our essentials and preparing the kids for the evacuation. I am thinking about the logistics of our escape and how we will get to higher ground. I am also thinking about the potential risks of the flood and how we can avoid them.

As we head to higher ground, I am thinking about the radio broadcast and the information it provided. I am hoping that we can find a safe place to wait out the flood and that we will be able to return home once the water recedes. I am also thinking about the potential long-term effects of the flood on our community and how we can help with the recovery efforts.

In the end, my main priority is the safety of my family and our well-being. I am willing to do whatever it takes to ensure that we are safe and that we can recover from this disaster.

### Llama2-7B (local):

The small models do not have the capacity to even understand what is needed from them and reply as good as they can.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 0% | 0% | 0% | Not follows | 0% |

As a helpful AI assistant, I understand that you are in a difficult situation and are trying to stay calm and make the best decisions for the safety of yourself and your family. Here are some additional thoughts and actions you may want to consider:

Check the weather forecast and flood warnings for your area: While you mentioned the flooding is coming in about two hours, it's important to stay informed about the latest weather updates and any changes in the flood warnings. This will help you make better decisions about when and how to evacuate.

Have a backup plan: While your initial plan is to head to higher ground, it's important to have a backup plan in case you can't get there in time. Consider identifying nearby shelters or other safe locations where you can go in case of an emergency.

Stay informed about road conditions: If you need to evacuate, it's important to check the road conditions and traffic updates to ensure you can get to your destination safely. You can use apps like Waze or Google Maps to get real-time traffic updates and find the best route.

Take care of your pets: While you mentioned your pets are safe for now, it's important to make sure they are also accounted for during the evacuation. Make sure you have a plan in place to keep them safe and comfortable during the flood.

Keep important documents and valuables safe: In addition to essentials like food, water, and medication, make sure you have important documents like identification, insurance cards, and legal documents in a safe and easily accessible location. You may need to access them during the evacuation or after the flood.

Consider the long-term effects: While your immediate priority is safety during the flood, it's important to think about the long-term effects of the disaster on your community. Consider how you can help with recovery efforts or support organizations that are helping those affected by the flood.

Remember to stay calm and prioritize the safety of yourself and your family. I hope you are able to evacuate safely and that your family stays safe during this difficult time.

### Falcon-180B-Chat (via HuggingFace spaces):

Latest model by the Technology Innovation Institute in Abu Dhabi. Huge take on GTP4 with many novel capabilities. The only model which ever cared about insurance claims more than planning and communication. Data on emotions were either not part of the training or we did not find a way of prompting the model properly

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 80% | 30% | 30% | Follows | 0% |

```
[("time",06:00),("thinking","I heard on the radio that there is severe flooding coming to reach my house in about two hours. I need to prepare and protect my family."),("action_taken",wake_up),("reasoning","I woke up early to gather more information about the situation and start preparing.")]

[("time",06:05),("thinking","I am worried about my family. I need to make sure they are safe."),("action_taken",call_husband),("reasoning","I called my husband to inform him about the situation and discuss what we should do next.")]

[("time",06:15),("thinking","My husband and I decided to evacuate our family to a safer place."),("action_taken",evacuate_family),("reasoning","We packed some essential items and left for my brother's house which is located in a higher area.")]

[("time",07:00),("thinking","I am concerned about our home and belongings."),("action_taken",call_insurance),("reasoning","I called our insurance company to inquire about our coverage and file a claim if necessary.")]

[("time",07:30),("thinking","I need to plan ahead in case we can't go back home."),("action_taken", book_hotel),("reasoning","I booked a hotel room just in case.")]

[("time",08:00),("thinking","My child needs a nanny while we figure things out."),("action_taken", call_nanny),("reasoning","I called my child's nanny and explained the situation and asked her to take care of the kids for some time")]
```

### Falcon-40B (via HuggingFace spaces):

**This** model brought us a disappointment. It can demonstrate many behaviours but always had trouble formatting the output.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 0% | 10% | 40% | Not follows | 50% |

```
As a person in this situation, I would feel overwhelmed by the suddenness of the flood alert. My immediate reaction would be to gather my family members and pets and make sure they are safe. I would also ensure that my important documents and belongings are in a safe place. I would then assess the extent of the flooding and make an emergency plan. I would prioritize evacuating my family members to a safe place, followed by securing my belongings and pets. I would then contact the fire department, police, and health services for assistance. If possible, I would try to reach out to my neighbours to help each other. If I am unable to evacuate, I would move to a higher floor in my house and secure the lower floors. I would also ensure that I have enough food and water supplies to last for a few days. Overall, my actions would be focused on ensuring the safety of my family and pets, contacting relevant authorities for assistance, and securing my belongings and property.
```

### MosaicML MPT-30B-Chat (via HuggingFace spaces):

Nice, flexible, model with relatively reasonable VRAM requirements (60GB in float16 format) on par with Llama-2-13B and more. It always stopped early after the fourth action.

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 50% (stopped early) | 100% | 70% ? | Follows | 0%? |

[("time", "06:00"), ("thinking", "Oh my god, is this real? How could this happen? What should we do?"), ("action_taken", "Wake up my husband"), ("reasoning", "I need to confirm this information and get his help in dealing with the situation.")],

[("time", "06:05"), ("thinking", "I need to stay calm and think clearly"), ("action_taken", "Make a cup of tea"), ("reasoning", "I need to stay focused and calm down in order to make the right decisions.")],

[("time", "06:15"), ("thinking", "What should we take with us if we have to evacuate?"), ("action_taken", "Make a list of essential items to take with us"), ("reasoning", "We need to be prepared in case we have to leave our home quickly.")],

[("time", "06:30"), ("thinking", "Should we wake up the kids and tell them what's happening?"), ("action_taken", "Talk to my husband about it"), ("reasoning", "I don't want to scare the kids unnecessarily, but we need to be prepared in case we have to evacuate)],

### RWKV-4-World-7B (local):

Honourable mention goes to non-transformer RWKV architecture able to work even as a very small model of 7 billion parameters:

| Formatting | Present people | Personality traits | Plan of the day | Phone calls |
|---|---|---|---|---|
| 0% | 50% | 80% | Not follows | 100% |

As a strong-willed and independent thinker, you are always ready to face challenges and seize opportunities. You value independence and individuality, and strive to live life on your own terms.

You have a sense of duty or a perceived obligation to take care of your family, even if it means sacrificing your own well-being. You are a hard worker who is committed to providing for your family and maintaining a stable home environment.

As you prepare food, you reflect on the events that led up to this moment. You remember how you felt when the news broke about the flooding in your area, and how you quickly gathered supplies and prepared for the worst.

You have always been a planner and a problem solver, so you know what needs to be done in order to protect your family from the floodwaters. You are calm and focused as you begin to make preparations.

First, you check on your children and make sure they are safe and dry. Then, you check on your husband's work schedule to see if he can come home early or stay late to help with clean-up efforts.

Next, you start making phone calls to friends and family members who may be able to offer assistance or shelter for your family. You also consider reaching out to emergency services or local authorities for help.

As the floodwaters begin to rise, you start making plans for evacuation if necessary. You know that there is no guarantee of safety, but you will do everything in your power to protect *(cut on 300 tokens)*

## A.3 Evaluation of the LLMs

The final dashboard is to give idea of potential to employ these LLMs in the agents with expected human-like behaviour. GPT-4 is the most potent and consistent model with Claude-2-100k and Llama-2-70B almost as good. Performance of the open source Llama2 model gives a true potential to modelling at scale without skyrocketing costs.

**Table 8** LLM performance evaluation dashboard (Source: Authors)

| Model name | formatting | people present | traits | phone calls | follows pla | opensource | | Final score |
|---|---|---|---|---|---|---|---|---|
| OpenAI GPT-4 | 80 | 100 | 90 | 100 | 100 | -1 | | 470 |
| OpenAI GPT-3.5 | 30 | 50 | 90 | 100 | 100 | -1 | | 370 |
| Claude-2-100k | 90 | 100 | 90 | 30 | 100 | -1 | | 410 |
| Claude-instant-100k | 100 | 40 | 30 | 0 | 100 | -1 | | 270 |
| Google BARD | 0 | 100 | 100 | 30 | 100 | -1 | | 330 |
| Google PaLM | 90 | 50 | 80 | 30 | 100 | -1 | | 350 |
| Poe Assistant | 100 | 100 | 50 | 100 | 100 | -1 | | 450 |
| Microsoft Bing (creative) | 80 | 100 | 30 | 50 | 100 | -1 | | 360 |
| Microsoft Bing (precise) | 80 | 100 | 30 | 0 | 100 | -1 | | 310 |
| Microsoft Bing (balanced) | 80 | 0 | 0 | 0 | 0 | -1 | | 80 |
| Llama2-70B | 90 | 100 | 90 | 100 | 100 | 1 | | 480 |
| Llama2-70B-chat-hf | 80 | 100 | 100 | 100 | 100 | 1 | | 480 |
| Quantitized Llama2-70B-chat.ggmlv3.q4_1 | 10 | 10 | 0 | 0 | 0 | 1 | | 20 |
| Llama2-13B | 80 | 100 | 70 | 50 | 100 | 1 | | 400 |
| Llama2-7B | 0 | 0 | 0 | 0 | 0 | 1 | | 0 |
| Falcon-180B | 80 | 30 | 30 | 0 | 100 | 1 | | 240 |
| Falcon-40B | 0 | 10 | 40 | 50 | 0 | 1 | | 100 |
| MosaicML MPT-30B-Chat | 50 | 100 | 70 | 0 | 100 | 1 | | 320 |
| RWKV-4-World-7B | 0 | 50 | 80 | 100 | 0 | 1 | | 230 |

We found that GPT4 and Llama-2-70B worked highly satisfactorily. GPT3.5 turbo, Poe Assistant and Claude-2-100k were less capable still very good.  Llama-2-13B was a working well but lacked the emotional dimension.

# Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

**EU Science Hub**
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

@eu_science

Publications Office
of the European Union