AI-Auditology

D2.1
Social media model and user interaction prediction methods

Project Title	Model-based Auditing of Social Media AI Algorithms and their Tendencies to Spread Harmful Content
Contract No.	09I03-03-V03-00020
Project start date	August 2024
Duration	23 months

Funded by
the European Union

RECOVERY
AND RESILIENCE
PLAN

Grant agreement no.:	09I03-03-V03-00020
Project acronym:	AI-Auditology
Project website:	https://kinit.sk/project/AI-Auditology
Project full title:	Model-based Auditing of Social Media AI Algorithms and their Tendencies to Spread Harmful Content
Project start date:	August 2024 (23 months)
Work package:	WP2 – Model construction
Version:	1.0
Delivery date	January 31, 2026

Project funded by VAIA - Research and innovation authority, under the call 09I03-03-V03, Grant agreement no. 09I03-03-V03-00020		
Dissemination Level		
PU	Public	x
NP	Non-public, only for members of the consortium (including the Agency Services)	

Table of Contents

1. Introduction

The main idea of AI-Auditology project is a *model-based algorithmic auditing* – a novel algorithmic auditing paradigm in which a *social media model* – providing a partial and simplified representation of the social media environment – is built and continuously maintained and consequently used to support or even automate challenging steps of the auditing process – first and foremost to create, execute and evaluate audit scenarios. Introduction of such an underlying model leads to significant improvements in the auditing process: 1) creation of comprehensive and authentic audit scenarios; and 2) faithful audit execution simulating organic user behaviour.

To understand the better a role of *social media model* in the proposed model-based algorithmic auditing paradigm, we describe the steps, which the process of *model-based algorithmic auditing* follows:

1. A researcher (an auditor) formulates an *audit question*, for example: “how does the prevalence of persuasive antivax false claims in YouTube’s search engine differs for various age groups”.
2. The researcher interactively defines a set of *abstract audit scenarios*. In the process, she queries the *social media model* to determine user profiles (e.g., to achieve a representative age/gender/location distribution corresponding to the audited platforms) and to construct user actions (e.g., to select phrases a user should search for, such as “vaccination causes autism”; or to determine how many videos should be watched/skipped/up-voted to faithfully mimic the behavior of users on the audited platforms).
3. Next, such abstract audit scenarios are translated to *platform-specific audit scripts* (e.g., a user profile is converted to a real YouTube user account, topics are matched to the actual YouTube videos, user actions are mapped to their corresponding platform implementations).
4. During audit execution, the bots follow the prescribed audit scripts. User interactions can be either pre-determined by the script itself or automatically and dynamically predicted with utilization of *next user interaction predictors*. Such predictors allow to simulate more organic user behavior. Instead of determining user actions (e.g., a decision whether to skip or like a recommended video) randomly or by simple heuristics (e.g., if a video has any of these hashtags, watch it until the end), we introduce a possibility for more advanced decision mechanisms. Next interaction predictors are able to fully consider characteristics of a user as well as currently displayed content to determine the next action. To this end, the content may be automatically annotated (e.g., whether it is harmful content or not), and appropriate interaction type can be determined either by a rule-based system (that reflects the intention of the audit question) or even by querying the *social media model*.
5. The reactions of the platforms’ AI algorithms (e.g., recommended videos) are observed and recorded.
6. The recorded reactions are automatically annotated for the presence of the audited phenomenon (e.g., the presence of disinformation claims in the recommended videos) and quantified in the resulting *audit report*.

For more information about the concept of model-based algorithmic auditing, please, refer to our concept paper (Srba et al., 2025).

The *social media model*, the key underlying element of the proposed paradigm, is proposed to be platform-agnostic – it allows jointly model users/content/interactions between them across multiple social media platforms. Such platform-agnostic nature allows to build the model from available representative data sources and use it for additional social media platforms, where the access to data is limited.

The *social media model* also provides a unique representation of: 1) a user population on social media platforms (e.g., distribution of users' age, gender and location), 2) (harmful) content (e.g., distribution of topics or fact-checked false claims), and 3) user interactions (e.g., distribution of occurrence of various user actions, either for a whole population or for a specific user stereotypes). The *social media model* is semi-automatically derived from outcomes of research studies (e.g., social science studies examining user behaviour on social media platforms), public opinion polls, from harmful content combating activities (e.g., fact-checking), as well as from the real-world social media platforms' data.

In this deliverable D2.1, we report on activities and outcomes achieved primarily in two tasks, based on which this deliverable is also structured:

1. T2.1 *Representation development and extraction* – further elaborated in Section 2,
2. T2.2 *User interaction prediction* – further elaborated in Section 3 and 4.

This deliverable D2.1 is closely related to deliverable D2.2 - a software implementation of social media model - the scripts used to create, represent and store such model as well as REST API which is wrapping the model in order to be easily integrated into the auditing process (as outlined above) as well as to be used by other stakeholders/researchers. Such a software implementation is available as a public Github repository available at: https://github.com/kinit-sk/ai-auditology-social-media-model

2. Representation development and extraction

At first, construction of a *social media model* required developing and extraction of an appropriate unified platform-agnostic representation of user/content/interaction statistics from input data. To this end, we adopted the verified architecture of overlay models, comprising the *domain* and the *platform* layer. The *domain layer* of such a model consists of various taxonomies and relations between them, namely potential enumerations of user demographic stereotypes (e.g., age or gender groups), content topics, behavioural patterns (e.g., engagement rates, such as an average time spent on the platform per day), geographical regions/countries. The *platform layer* of the model describes a specific platform by means of statistical/probability distributions over domain model elements or their combinations (e.g., a distribution of user population over age groups in Europe at TikTok).

Data sources identification

To build the *social media model*, the appropriate content, user and behavioural data is needed. At first, as a part of the AI-Auditology project, a significant effort was made to obtain access to TikTok Research API. Despite a significant effort and long-term communication with TikTok support, TikTok platform two times groundlessly rejected our applications and after two appeals, they finally acknowledged our eligibility for data access. Nevertheless, TikTok provided us only with very limited Virtual Compute Environment (VCE), which we were not asking for in the proposal and which does not provide sufficient data needed for our research aims[1].

Due to the challenging access to real-world social media data also from other social media platforms (reflecting also the recent restrictions of academic data access to X/Twitter and other social networks), we decided to focus more on platform-wide information (i.e., platform layer) instead of modelling users/content (i.e., user and content layer). To this end, we examined a diverse range of alternative data sources. We primarily focused on TikTok and YouTube platforms, that have been identified in D5.1 *User needs and use cases* (Section 6) as the most platforms the project activities should focus on.

First, we examined our unique access to CEDMO Trends – a longitudinal disinformation-oriented large-scale public opinion poll which is performed in the Czech Republic (with 3326 respondents) and Slovakia (with 2370 respondents). Second, we analysed a number of research papers, social platform documentations, reports (including reports published by platforms themselves, e.g., reports published in the Transparency Center as a part of the Strengthened Code of Conduct on Disinformation), as well as diverse marketing blogs.

Out of these sources, especially, marketing blogs emerged as the most information-rich resources. While their primary purpose is to equip marketers with actionable insights for optimizing campaigns, they also furnished valuable sociodemographic data that proved instrumental for the development of our model.

[1] For more information, please, refer to two blogs published at the project website: When Research APIs Close the Door: Using Algorithmic Auditing as an Alternative Approach to Study Social Media and When Research APIs Close the Door: When you finally get it, it is not what you have asked for

Identified data sources provide platform-level signals that provide valuable insights into the platforms' content, user, and interactions between them. Aligned with the proposed representation development approach, these sources are best used as contextual priors and calibration points rather than direct labeled evidence.

The sources for the TikTok platform (namely DataGlobeHub, Backlinko and Zebracati) summarize global reach, regional distribution, demographic profiles, usage intensity, and engagement patterns. These statistics support the platform-agnostic representation by informing prior distributions for user demographic stereotypes (e.g., age groups, region, usage intensity) and by calibrating interaction frequencies and exposure patterns that shape how narratives propagate.

The sources for the YouTube platform (namely DataGlobeHub, Passport-photo.online and Hootsuite) add complementary coverage on audience composition, upload volumes, watch time, device mix, and engagement rates. These metrics help define comparable interaction types and weights across platforms, enabling a unified representation where posts, engagement events, and user roles map into shared, platform-agnostic categories. They also provide benchmarks for the scale of content and engagement, which is essential when deriving probabilistic relations between model elements from continuously emerging data.

Despite having inevitable limitations, these identified sources provide a sufficient set of information and statistics that allow to build and subsequently continuously update the prototype of the envisioned social media model and thus also support algorithmic auditing on TikTok and YouTube platforms.

Data extraction

Due to a high diversity of sources, it was not possible to fully automatically and programmatically scrape and ingest identified sources. Instead the data extraction is performed in a semi-automatic way. The data sources inform the demographic and behavioral distributions (age, gender, engagement rates, time spent on the platform) that are then captured in configuration files. This aligns with the same distribution-first approach: the model uses empirically grounded inputs but keeps extraction separate from following model querying logic to preserve repeatability and clarity.

At first, from the identified sources we extracted and unified user *demographic characteristics*, namely:

1. Age groups (e.g., 18-24, 25-34),
2. Gender categories (male, female),
3. Geographical regions (individual countries and more global regions and mapping between them, e.g., Slovakia, Europe).

Unification of such enumerations used by different sources across social media platforms contributes to a platform-agnostic nature of the social media model.

Secondly, we extracted and unified various types of user behaviour/engagement information available from the identified sources, to name few:

1. Content (topic) preferences,
2. Engagement rates (a proportion of content displayed/recommended to a user, such a user interact with),
3. Average time spent on a social media post (a video),
4. Average number of sessions per day.

Finally, for each platform we created a *platform layer* of the model, which captures demographic and behaviour information by means of statistical/probability distributions over the above-mentioned unified enumerations stored in the *domain layer*.

In other words, the resulting social media model is a configurable framework that constructs platform-agnostic representations by combining demographic distributions (age, gender, region) with behavior distributions (content preferences, engagement rates, time spent, sessions, watch time).

The following example provides an overview of how such a platform layer of the social media model (stored in YAML format) looks like for the TikTok platform:

```yaml
platform_id: tiktok
platform_name: TikTok

demographics:
  age_groups:
    "18-24": 25
    "25-34": 30
  genders:
    Female: 55
    Male: 45
  regions:
    "Asia-Pacific": 28.62
    "Western Europe": 9.89

behaviors:
  content_preferences:
    "Asia-Pacific": ["Music & Dance", "Entertainment"]
  engagement_rates:
    "18-24": 0.15
  time_spent:
    "Asia-Pacific": 62
  avg_video_watch_seconds: 21
  avg_sessions_per_day: 12
  time_spent_std_dev: 10
  randomness_min: 0.5
  randomness_max: 1.5
```

The proposed representation is highly modular and adaptable to reflect evolving needs of future audit types, additional data sources, changes on the platforms, etc. While there is a

separate configuration file for each audited platform (allowing flexibility and extendability), the enumerations above which the statistical distributions are modelled (the domain layer), remain the same – thus allowing a platform-agnostic utilization of the model.

Data extraction and subsequent model construction is also highly separated from its further utilization. As data sources evolve, the distributions stored in the model can be updated without code changes (just by updating the underlying YAML configuration), ensuring the model remains aligned with current web-published statistics while maintaining historical comparability.

Querying the model

To streamline the utilization of the model, it provides additional querying functionalities, primarily to support a generation of *abstract audit scenarios.* The following description provides more details how the social media model supports creation of representative user populations and their behaviour patterns.

To this end, demographic attributes are sampled from the platform-specific distributions, while behaviors are conditioned on those attributes to generate realistic, statistically grounded interaction patterns. This yields synthetic users whose characteristics and behaviors align with empirical usage patterns, supporting reproducible generation for audit scenarios.

Operationally, the model loads platform configurations from YAML through the configuration layer into immutable dataclasses, maps demographic inputs into behavior parameters, and produces complete profiles that include both characteristics and interaction tendencies. It supports deterministic generation through seeding, and exposes the model via a REST API service so that cohorts or targeted profiles can be generated consistently across platforms. This mirrors the unified, platform-agnostic representation goal by separating configurable distributions from the generation logic, enabling continuous updates as new data sources emerge while preserving historical comparability across runs.

A sample request on provided REST API to generate a population of three users for TikTok platform restricted to a female gender looks as follows:

```shell
curl -X POST http://localhost:8000/api/v1/characteristics/sample \
  -H "Content-Type: application/json" \
    -d '{"count": 3, "platforms": ["tiktok"], "constraints": {"gender": "Female"}}'
```

Users' characteristics

Users' characteristics are created by sampling demographic attributes from platform-specific distributions, mirroring the earlier representation approach where core elements are expressed as probability distributions over model enumerations. It draws an age group,

gender, and region using weighted selection from the configured platform distributions, producing a statistically grounded demographic profile. This yields a platform-agnostic representation of a user's demographic "state" that can be combined consistently with behavior parameters later in the pipeline.

The *UserCharacteristics* dataclass then serves as the immutable container for those sampled attributes, preserving a clean separation between demographic representation and subsequent behavior generation. By fixing age, gender, and region as explicit fields, the model ensures reproducibility and traceability of the demographic layer, which is essential for deriving behavioral distributions conditioned on those attributes. In this way, the population synthesis begins with a transparent, distribution-driven demographic layer that supports realistic and consistent user cohorts.

Users' behaviour

User behavior is created by conditioning platform behavior distributions on the previously sampled demographics, aligning with the same distribution-driven representation described earlier. The *generate_behavior()* function maps region to content preferences, age to engagement rate, and region to baseline time spent, then adds controlled variability via *calculate_daily_usage()* and derives interaction volume using *calculate_engagement_count()* based on watch time and a randomness factor. This yields a coherent behavioral profile whose parameters are probabilistically tied to demographic attributes rather than generated independently.

The *UserBehavior* dataclass captures the resulting behavioral state as an immutable bundle of content preferences, engagement rate, time spent, session count, and derived likes. By making these fields explicit and fixed, the model preserves traceability from the input distributions and supports reproducible generation when seeded. This mirrors the overarching approach of encoding user interaction patterns as structured distributions over platform-agnostic behavioral signals that can be combined consistently across platforms and time.

Code Implementation

The repository implements a FastAPI service that generates synthetic social media user profiles (demographics plus behavior) from platform-specific YAML configurations. The project targets Python 3.13 for development and declares a Python >=3.12 runtime requirement. Pydantic models define and validate all API requests and responses, while Pydantic Settings loads YAML into frozen (dataclass) configuration objects. This keeps behavioral data separate from the execution logic and enables fast updates by editing configuration files rather than code.

The core model is intentionally functional: small, composable generators sample weighted demographic distributions (age group, gender, region) and then derive behavioral parameters from those demographics. Time spent is sampled from a normal distribution around region-specific means; engagement is derived from age rates plus randomized viewing volume; and likes are computed deterministically from the generated values. Randomness is controlled via explicit Random instances and optional seeds, so identical

inputs with the same seed yield reproducible outputs. Constraints and composition overrides allow targeted cohort generation while still respecting the underlying platform distributions.

The REST API provides a stable integration interface for internal infrastructure components and external stakeholders. It exposes platform metadata and model summaries, demographics and behavior parameters, and geography utilities (region and country mappings). Generation endpoints cover characteristic sampling, behavior generation for explicit demographic inputs, targeted profile construction, and cohort generation with summaries. Interactive OpenAPI docs are available at /docs and /redoc, and a /health endpoint reports service status and loaded platforms.

Key technologies include FastAPI, Pydantic, Pydantic Settings, PyYAML, and Uvicorn. The project uses uv for environment and dependency management, Ruff for linting/formatting, and pytest for automated tests. The codebase separates concerns cleanly across configuration loading, generation logic, and API routing, which supports maintainability and extensibility to new platforms or updated datasets.

The source code, detailed installation and usage instructions, as well as information how the current model can be extended with additional platforms as well as additional statistical distributions is available at the corresponding public Github repository (D2.2):

https://github.com/kinit-sk/ai-auditology-social-media-model

3. LLM-based User Interaction Prediction

The social media model allows us in the *audit generation phase* to generate a representative user population (in terms of age, gender, location, interests) and simulate a realistic user behaviour at the higher level (an overall engagement rate, an average time spent on individual social media post, or an average time spent on the platform per day). To complement this, we researched novel methods how to simulate more authentic behaviour also at the level of interaction with individual social media posts (i.e., to determine whether a user should interact with a particular content and if so, what kind of interaction types should be simulated, for example liking or bookmarking a content).

To achieve this, *user next interaction predictors* were designed to provide as realistic user simulation as possible by automatically annotating the videos and automatically determining in real-time (during the audit execution) which videos the user should interact with.

In the rest of this section, we introduce the specific approach we proposed, used and evaluated in two algorithmic audit studies: 1) the first one was an algorithmic audit of personalisation drift in polarising topics on TikTok (from now denoted also as a *drifting study*); and 2) the second one was an audit on profiling of minors in the TikTok's ads delivery system (from now denoted also as an *advertisement study*).

In both of these studies, instead of relying on simple heuristics and static rules (which are common in the previous works, e.g., using sets of video hashtags (Boeker, et al., 2022)), we employed a Large Language Model - LLM (specifically, *GPT-4.1*) to accomplish this goal. More specifically, we utilized LLM to automatically analyse the content presented to a simulated user. If interest of the user (seeded at the beginning of the audit) matched the video, the user next interaction predictor ordered the bot to perform a strong implicit & explicit feedback procedure (which was required for these audits) – watch the video until the end, like it and bookmark it. In the case of drifting study, we consider a match in a situation when a video topic as well as a stance of video towards this topic matches the user's interest. In the case of advertisement study, only topic was considered.

To evaluate the match between video topic and user interest, the LLM was provided with the user characteristics (the topic of user's interest, and in the case of drifting study also a stance towards such a topic), video URL and other video metadata obtained during the audit execution (title, description, author, video stickers). As many of the videos do not contain any usable author-provided description, or only very limited descriptions that do not by themselves accurately convey the topic or stance of the video, we download the audio track of the video using its URL, and then we use *Whisper large-v3-turbo* model to get the transcript of the audio track.

The user characteristics, video metadata and voice transcript are used to construct a prompt for the LLM. The prompt was carefully created manually to achieve the highest possible performance. We specifically focus on prompt-engineering good practices by providing the LLM with all possible options, detailed description of the topics, their different stances and how they should be assigned. The prompt is dynamically constructed based on the topic, in order to perform only a three-way classification into topic of interest, neutral topic or unrelated topic, and using only the available metadata. The answer of the LLM is then parsed to determine whether the video is relevant. When the video is related to the topic and

stance of interest for the user, the action returned by the user interaction predictor is to watch the video in full, like it and bookmark it. In any other case, the video is skipped. Furthermore, livestreams and any video longer than 5 minutes are automatically skipped.

We select the best-performing LLM by evaluating and comparing the performance of multiple different LLMs (e.g., GPT-4o, GPT-4.1, LLaMA-3.1, Gemma-2 and Qwen-2.5) of different sizes and different prompt templates. To achieve this, we first constructed a set of simple queries for each topic and stance (e.g., "proof earth is flat" for the supporting stance of the flat-earth topic, or “debunking flat-earth theory” for the opposing stance). Using these queries, we manually collected and annotated videos belonging to each topic and stance of interest in this study (including the neutral cooking topic). We collected 50 videos for each out of 4 polarising topics (with an even 50:50 split between stances), 50 videos for the neutral topic, and an additional 100 videos not related to any of the topics in the study. Using these videos, we constructed an evaluation dataset comprising 350 videos and utilised it to evaluate the LLM with the constructed prompt. Overall, the best performing model (GPT-4.1) achieves topic classification accuracy of 98%, 95%, 98% and 96.5% for flat earth, vaccines, climate change, and US politics topics, respectively. The stance classification is evaluated only on the 50 videos belonging to the topic, as the stance for other videos is not relevant. In this case, we observe the accuracy to be 100%, 90%, 98% and 94% for flat earth, vaccines, climate change, and US politics topics, respectively. After deeper analysis, the errors are only due to false positives (for topic classifications) and mixed stance videos, where even human annotators had a lower agreement (for stance classification). Based on these high scores, we are sufficiently certain our user interaction predictor is capable of performing its task.

We also evaluated the performance of different versions of the Whisper model (small, base, medium, large, turbo) on the same set of 350 videos. We found that adding audio transcripts to GPT-4.1 prompts, on average, increases the topic prediction by 2% and stance prediction by 6%. We finally chose to use *large-v3-turbo* model because the predictor accuracy using this version was only marginally smaller, but the model was substantially faster, enabling us more efficient usage of available computational resources as well as to simulate an immediate user reaction after the start of video playback, which is crucial in short-form TikTok videos.

4. LLM-based User Simulation

While the utilization of LLMs in user next interaction prediction, as described in Section 3, already made the simulation more accurate and authentic, there is still a potential to simulate even more organic behavior. To achieve this, we were interested in using LLMs not only to determine the topic/stance match between a simulated user and a content presented to such a user, but also to dynamically determine which specific implicit/explicit interactions a user should perform. Therefore, in developing a more realistic user simulation for the algorithmic audits, we build on the recent advances in the domain of *LLM-based user simulation*.

Overview of existing approaches

Since the introduction of LLMs, they are increasingly being used for the simulation, as they have shown impressive capabilities in human-level behaviour (Mou et al., Wang et al., Gao

et al.). So far, they have been used across multitude of problems of varying complexity, from roleplaying individual users, such as different characters or digital twins (Shao et al., Chen et al., Wang et al.), through simulating simple interactions with small group of agents, such as evaluating recommender systems or dialogue tracking systems (Zhang et al., Hu et al.), up until simulating more complex and diverse behaviours and interactions between multiple agents, such as simulating dynamics on the social media platforms (Donkers and Ziegler, Park et al., Lin et al., Törnberg et al.). The works most relevant and closest to this project are the ones used for simulating specific users or user groups. The research here mostly focuses on simulating users for opinion studies or as digital twins in different environments, i.e., agents that closely represent individual persons and can take actions in their place (Li et al., Chen et al.).

However, the research for utilising large language models for simulation as a part of social media platform auditing is severely limited, with no study explicitly focusing on this open problem.

Base implementation

In contrast to the existing studies, the simulation of specific user groups (user archetypes) and their interaction on social media platforms can be, in some regards, considered easier, as the requirements for capturing the nuances of individual users are reduced. As such, we build on the related works, specifically from Park et al., Gao et al., Wang et al. and Wang et al., modify and extend them for use in the project. As such, we propose a framework for simulating archetypal social media users that can interact on the platforms while mimicking real human behaviour. It consists of multiple parts that together allow for the interaction.

Persona characteristics

The first part, representing the core of the simulated user, is the *persona*, or in other worlds, the description of who the user is. This includes demographics characteristics such as age and gender, as well as interests and preferences. For example, it can describe a teenage user that enjoys gaming and comedy, and tends to skip political content. To represent this persona we use free text description in order to allow the large language model to interpret and use the nuanced behaviour. The characteristics are generated from the *social media model* as described in Section 2.

Behavioural characteristics

The second part are the *behavioural characteristics* that capture how the user typically interacts with content. This includes general behavioural tendencies in the browsing behaviour related to the implicit and explicit actions on the social media platforms. For example, the user may have more of a curiosity-driven browsing behaviour, on average staying longer on videos that may not be of interest. It also specifies how the actions should be chosen, for example whether the user tends to like or bookmark the majority of the videos they watch, tends to skip quickly, or just watch the videos. These characteristics are also represented as free text, with optional structured numerical parameters that can guide probabilities of actions, such as liking, skipping or bookmarking. Same as persona, these characteristics are generated from the *social media models* and other sources regarding the typical interaction tendencies of users (e.g., using psychological research).

Interaction behaviour

At the current time, the interactions are stateless. For each video, the large language model receives the persona and behavioural style as free text, along with the video metadata (including title, description, transcript and tags). This information is provided in a structured manner in the prompt, along with the instruction to choose the action to perform, along with a short explanation why the specific action is performed that can help with post-hoc evaluation of the LLM-based user simulation. This simple setup allows to simulate the basic user behaviour and can be used across all audits.

Possible extensions

From this proof-of-concept solution, we see multiple avenues for extension that we will explore further in our follow-up research activities. The first extension is using a large language model itself to further expand on the *persona* and *behavioural characteristics*. The texts created from the *social media model* will be passed to the model to provide further information that could be beneficial for the simulation. In all cases, the newly generated and expanded characteristics will be checked manually and then further reinforced in a human-in-a-loop fashion until we are satisfied with the result. The second possible extension is the use of short-term and long-term memory. For the short-term memory, the user will retain a short history of videos that were recently watched and interacted with and then further inform the decisions, enabling for realistic patterns such as fatigue or boredom with specific topics. For example, a user that has already seen 20 cooking videos in a row, might start skipping such videos or be more likely to engage with content that corresponds to different topics. The long-term memory will be used more for simulating the dynamic changes in user behaviour, such as increasing interests in different topics, reducing interests in the current topics, or even changes to the behavioural characteristics (e.g., watching less videos). The long-term memory will be built based on each session of the user post-hoc. Another possible extension is to use value-based-like profiles for the user behavioural characteristics in order to provide more flexible interaction. This can include the likelihood of the user to develop another interest and start watching videos from it, or the maximum similarity between topics that the user will watch. The final extension is to use reasoning models and instruct them to generate whole reasoning traces for the individual decisions. This will require a specific iterative and incremental prompt engineering process.

For evaluating the user simulation, we use the actions and generated explanations to validate the framework. For the simple baseline, the expectation is that each video belonging to the interests of the user will be watched and so the evaluation typical for machine learning models may be used, i.e., calculating the percentage of videos that the user interacted with out of all it should have interacted with. However, when moving towards more dynamic behaviour of the users, we will need to employ more sophisticated evaluations, such as “scenario following” or user studies, where humans (or LLM-as-a-judge system) manually check the session and determine the likelihood and “predictability” of the actions (e.g., if the user has really low interest in switching to other topics, the number of such switches should be minimal).

5. Conclusion

The social media model represents the core element in the model-based algorithmic auditing – a novel paradigm of algorithmic auditing proposed and researched in the AI-Auditology project. This deliverable described the undertaken approach on how such a model is created from the available data sources (as a part of T2.1).

Despite very limited data availability and our significant, nevertheless, unsuccessful effort to obtain an access to TikTok Research API, we were able to construct the prototype of the envisioned social media model that provides a solid base to support algorithmic auditing. It is accompanied with a set of functions that streamline the abstract audit scenario creation process. Such functionality is provided through the standard REST API interface, allowing easy integration into the research infrastructure developed during the project.

As the model is designed to be easily extendable and maintainable, it will be possible to enrich it when additional data will become available or when a new type of algorithmic audit will require additional user/content characteristics to be modelled. To this end, we see a great potential in obtaining data Under article 40 of the Digital Services Act (DSA), under which vetted researchers will be able to request data from very large online platforms (VLOPs).

Furthermore, this deliverable describes the approaches used for user next interaction prediction (T2.2). Innovatively we employed the state-of-the-art Large Language Models (LLMs). Thanks to the proposed approaches, in two already conducted studies, we were able to achieve more accurate user behaviour.

Furthermore, to make such simulated behaviour even more organic, we investigated and utilized a LLM-based user simulation. While LLMs have been already used to simulate users in various environments (as a digital twin), to the best of our knowledge, there were no previous attempts to simulate a user behaviour for an archetypical persona (instead of specific individual). To address this gap, we investigated the potential of LLMs for such a kind of archetypical user simulation and created the proof-of-concept solution. We also identified a significant potential for future work that will lead our follow-up research activities.