PersonaTrace: Synthesizing Realistic Digital Footprints with LLM Agents

Minjia Wang[1, 2], Yunfeng Wang[1], Xiao Ma[1], Dexin Lv[1], Qifan Guo[1], Lynn Zheng[1], Benliang Wang[1], Lei Wang[1], Jiannan Li[1], Yongwei Xing[1], David Xu[1], Zheng Sun[1]

[1]Apple, [2]Harvard University

Correspondence: minjiawang@g.harvard.edu

Abstract

Digital footprints—records of individuals' interactions with digital systems—are essential for studying behavior, developing personalized applications, and training machine learning models. However, research in this area is often hindered by the scarcity of diverse and accessible data. To address this limitation, we propose a novel method for synthesizing realistic digital footprints using large language model (LLM) agents. Starting from a structured user profile, our approach generates diverse and plausible sequences of user events, ultimately producing corresponding digital artifacts such as emails, messages, calendar entries, reminders, etc. Intrinsic evaluation results demonstrate that the generated dataset is more diverse and realistic than existing baselines. Moreover, models fine-tuned on our synthetic data outperform those trained on other synthetic datasets when evaluated on real-world out-of-distribution tasks.

1 Introduction

Digital footprints are the persistent records that individuals leave behind when they interact with digital systems — email threads, chat logs, calendar appointments, purchase histories, sensor traces, and more (Shiells et al., 2022; Kolawole and Rahmon, 2024).

Such traces fuel a wide range of downstream applications: they enable fine-grained user modeling and personalization (Valanarasu, 2021; Vullam et al., 2023), support behavioral and social-science research (Golder and Macy, 2014; Padricelli and Coppola, 2024), and provide large-scale supervision for data-hungry machine-learning pipelines (Zhao et al., 2023).

Unfortunately, progress in this area is throttled by data scarcity. Publicly available corpora cover only slivers of human activity. For example, the Enron email corpus (Klimt and Yang, 2004) captures a single company from the early 2000s. They also tend to focus on a single bundle — emails (Mehdi Gholampour and Verma, 2023; Greco et al., 2024), chat dialogs (Zhang et al., 2018; Suresh et al., 2024; Jandaghi et al., 2024), transaction logs, and so forth — failing to reflect the breadth of modern digital life. Proprietary data are subject to restrictive licenses, as raw digital footprints contain highly sensitive personal information. Regulations such as GDPR prohibit most forms of data sharing, and even internal access is tightly controlled. Anonymization alone is insufficient because rich textual artifacts can be deanonymized with modern LLMs (Panda et al., 2024).

Synthetic data generation offers a promising workaround and has demonstrated success in training state-of-the-art LLMs (Grattafiori et al., 2024; Qwen et al., 2025) and addressing tasks such as mathematics (Yu et al., 2023), coding (Wei et al., 2023), and general instruction following (Xu et al., 2023; Li et al., 2024b). Current synthesis methods, however, presume access to large seed dataset to bootstrap diversity — an assumption that breaks for digital-footprint data, where both public and private sources are largely inaccessible.

To address these challenges, we introduce *PersonaTrace*, a framework that synthesizes realistic, multi-bundle digital footprints with the help of LLM agents. PersonaTrace first creates persona profiles from a pre-defined demographical distribution. Given a profile, PersonaTrace simulates a plausible sequence of everyday events (e.g., attending a conference, shopping online, planning a family trip) and then generates the concrete digital artifacts that those events would leave behind (emails, SMS exchanges, calendar entries, reminders, etc).

We assess PersonaTrace with intrinsic metrics that quantify diversity and realism, and with extrinsic metrics that measure downstream utility. Specifically, we fine-tune open-source LLMs on the synthetic corpus and evaluate generalization on four real-world, out-of-distribution benchmarks: email

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 60–77

categorization, email drafting, question answering, and next-message prediction. Across tasks, models trained on PersonaTrace achieve competitive or superior results, compared to those trained on the strongest prior synthetic datasets.

Our contributions are as follows:

- We present the first end-to-end method for synthesizing *complete digital footprints* through a persona-driven workflow that ensures coherence and realism across user behaviors.
- Our comprehensive evaluation demonstrates that our dataset excels in both intrinsic properties such as diversity and realism, and extrinsic performance on downstream tasks.
- We release *PersonaTrace*, a high-fidelity synthetic digital footprint dataset and accompanying framework, to facilitate responsible future research [1].

2 Related Work

A key strategy for improving quality and diversity for LLM-generated synthetic texts is to guide generation using different priors.

Seed–dataset priors. Several approaches generate synthetic data by expanding seed datasets, such as conversational corpora (Jandaghi et al., 2024), instruction-tuning training sets (Xu et al., 2023; Huang et al., 2024; Li et al., 2024b; Gandhi et al., 2024), or domain-specific problem sets (Yu et al., 2023; Wei et al., 2023; Braga et al., 2024; Huang et al., 2025; Khalil et al., 2025). However, these methods are less suited for synthesizing digital footprints due to the lack of comprehensive seed datasets that span multiple modalities (e.g., emails, messages). Publicly available datasets typically focus on a single modality (Klimt and Yang, 2004; Zhang et al., 2021; Li et al., 2017; Chee et al., 2025), making it difficult to construct coherent multi-modal user profiles.

LLM-only priors. Some approaches rely solely on the generative capabilities of aligned LLMs without using external seed data (Xu et al., 2024). While attractive for open-weight checkpoints, commercial APIs typically forbid blank turns or control tokens, limiting its applications.

Intermediate-attribute priors. Some methods guide data generation using intermediate attributes such as knowledge taxonomies or persona descriptions (Li et al., 2024a; Ge et al., 2025; Tang et al., 2024; Fröhling et al., 2024). However, existing persona-based priors often emphasize professional or academic traits, resulting in biased distributions that do not reflect real-world user profiles and are unsuitable for generating realistic digital footprints.

3 Methods

Our approach employs an agent-based architecture built on LLM agents to simulate a realistic user and their digital footprint. We design a three-stage pipeline with specialized autonomous agents that collaborate to generate synthetic data. First, a Persona Agent constructs a detailed persona profile from an initial user specification. Next, an Event Agent expands this persona into a timeline of plausible events tailored to the persona's life. Finally, an Artifact Generator Agent produces diverse digital artifacts (e.g., emails, text messages, calendar entries, reminders, wallet passes) corresponding to these events, with Critic Agents iteratively reviewing the artifacts for consistency and realism. All agents share the same underlying LLM (Gemini-1.5-Pro with a temperature of 0.9) but are prompted with role-specific instructions and constraints. Figure 1 shows the overview of our proposed framework. Appendix C highlights prompts for each agent.

3.1 Persona Generation

In the first stage, the Persona Agent synthesizes a rich personal profile for the fictional user. We begin by sampling a set of basic demographic attributes from a predefined prior distribution (covering age, gender, locale, etc.) to ground the persona in realism. The priors are estimated from the 2022 American Community Survey (U.S. Census Bureau, 2022) to ensure plausible macro-level distributions. Given these attributes, the Persona Agent composes a basic identity for the user. This identity includes details such as name, age, gender, birth date, ethnicity, income level, household setup, and location. The Persona Agent then expands this profile by generating a social network context. It creates a list of key relationships (e.g., family members, close friends, coworkers) and assigns each connection basic characteristics (names, ages, relationship to the persona, etc.). To further enhance realism, the Persona Agent outlines the persona's typical daily routines and major life events. It provides a weekday routine and a weekend routine that reflect the persona's lifestyle. It also enumerates significant life events such as holidays, vacations,

[1] Data and code will be made available on request due to privacy and legal concerns.

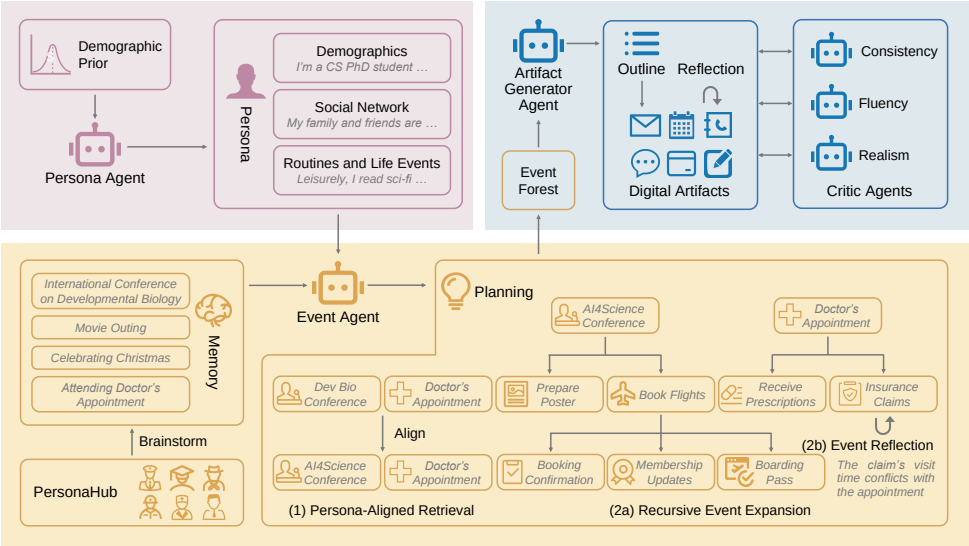

Figure 1: An overview of our methods. The **Persona Agent** generates a basic profile from a demographic prior, and iteratively adding realistic attributes to it. The **Event Agent** retrieves seed events from the event memory and aligns them to the persona, and brainstorms with self-reflection to generate an event forest that serves as the scaffolding of the digital footprints. The **Artifact Generator Agent** and **Critic Agents** forms a loop - the Artifact Generator Agent generates the outline and then digital artifacts, and the critic agents provides feedback to iteratively improve the quality of the artifacts.

and personal milestones. The outcome of this stage is a comprehensive persona profile that will inform subsequent event generation.

3.2 Event Generation

Event memory. We equip the Event Agent with an internal event memory $\mathcal{M}$ — a knowledge base that stores concise descriptions of activities people experience. To populate $\mathcal{M}$, we begin with PersonaHub (Ge et al., 2025) as the foundational source, which consists of descriptions of various personas. We then ask the Event Agent to brainstorm plausible daily-life events that a persona in PersonaHub might encounter. The combined list is then pruned for near-duplicates with MinHash LSH (Broder et al., 1998), yielding a diverse yet compact collection that serves as the agent's prior knowledge of the world.

Persona-aligned retrieval. Given a persona profile π, the Event Agent retrieves a subset of seed events from its memory $\mathcal{M}$: it selects the 30 most semantically relevant entries via embedding search[2], samples 30 uniformly for diversity, and synthesizes 40 fresh event prompts directly from π. The Event Agent then aligns each chosen seed event with the persona's details, modifying event descriptions as needed to fit the persona. For example, "attend an international conference on developmental biology" $\rightarrow$ "attend an academic conference on AI for science" if π describes a CS PhD student.

Recursive event expansion. The Event Agent expands each aligned seed event into a tree of sub-events, thereby constructing an event forest $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ for the persona. Each seed event serves as a root node in an event tree $\mathcal{T}$, and the Event Agent autonomously decides whether and how to branch that event into finer-grained sub-events. Some events may remain atomic, while others unfold into a sequence of related sub-events. For instance, a travel-related root event like "attend an academic conference on AI for science" might be expanded into sub-events such as "prepare poster" and "book flights". Then, "book flights" can be further expanded into "receive booking confirmation", "get membership updates" and "receive boarding pass". The Event Agent generates these sub-events with appropriate details and temporal ordering, effectively narrating how the larger event plays out. The Event Agent iterates breadth-first until (1) no more sub-events are expanded or (2) $|\mathcal{F}|$ exceeds 300 nodes, whichever comes first.

Event reflection. Moreover, the Event Agent reflects on each expanded event to ensure quality and completeness. It verifies that the sub-events provide sufficient detail, follow a logical structure,

[2]Embeddings are obtained from all-MiniLM-L6-v2.

	Dataset	Pairwise Corr. (↓)	Remote-Clique (↑)	Entropy (↑)	Avg. #Links	Avg. Length
Synthetic	FinePersonas-Email	0.2333	0.7680	2.8080	0.0000	681.50
	IWSPA-2023-Adversarial	0.4079	0.5919	2.5094	0.0000	276.48
	LLM-Gen Phishing	0.3094	0.6906	2.6969	0.2522	685.58
	Synthetic-Satellite-Emails	0.5416	0.4586	2.8218	0.0000	1050.76
	PersonaTrace (Proposed)	**0.2093**	**0.7898**	**2.8305**	0.2532	1437.87
Real	Enron	0.2798	0.7218	2.7110	0.0000	2002.57
	Human-Gen Phishing	0.1686	0.8334	2.9257	0.0020	3332.91
	Private	0.2066	0.7957	2.8332	13.6970	10036.44
	Private w/o Spam	0.2094	0.7893	2.7079	7.6918	5814.34
	W3C-Emails	0.1796	0.8196	2.7872	0.0000	2224.89

Table 1: Diversity and realism of datasets related to emails. Best results in synthetic datasets are in bold.

	Dataset	Tone	Fluency	Coherence	Informativeness	Engagement	Overall
Synthetic	FinePersonas-Email	4.52	4.92	4.56	3.98	3.98	4.39
	IWSPA-2023-Adversarial	1.59	1.91	1.67	1.31	1.22	1.53
	LLM-GenPhishing	3.41	4.89	4.67	3.13	2.78	3.70
	Synthetic-Satellite-Emails	4.82	4.96	4.86	4.91	3.65	4.64
	PersonaTrace (Proposed)	**4.95**	**4.99**	**4.99**	**4.92**	**4.09**	**4.79**
Real	Enron	4.19	4.73	4.47	4.28	2.71	4.07
	Human-Gen Phishing	3.50	4.06	3.69	3.73	2.40	3.47
	W3C-Emails	3.99	4.56	4.31	4.24	2.82	3.98

Table 2: LLM-As-Judge scores of datasets related to emails. Best results in synthetic datasets are in bold.

and are likely to result in digital records in the next stage. If any branch is found lacking (e.g., inconsistent), the Event Agent revises the event accordingly. The result of this stage is a collection of event trees – an event forest – that captures a diverse, personalized sequence of events the persona will undergo. This event forest serves as the scaffold for generating digital artifacts in the final stage.

3.3 Digital Artifact Generation

In the final stage, the Artifact Generator Agent and the Critic Agents work in tandem to produce high-quality digital artifacts for each event in the event forest. We adopt a generator–critic framework a generator–critic loop à la Madaan et al. (2023). For a given event ε and the persona context π, the Artifact Generator Agent first creates an outline of the digital artifact a — for example, an email, a text message, a calendar invitation, a reminder, a wallet pass, or other relevant digital record that would reflect what the persona would actually produce or receive in the scenario (e.g., an email confirming a flight booking or a text message exchange with a friend about an upcoming dinner). It then instantiates the artifact based on the outline. Once the Artifact Generator Agent proposes an artifact a, each Critic Agent evaluates it and provides feedback. Three Critic Agents evaluate the artifact for consistency with ε and π, as well as for its realism and fluency. They ensure that the content aligns with the persona's known attributes and the details of the event, flagging any contradictions, unnatural language. After receiving the critique, the Artifact Generator Agent revises the artifact a accordingly. This generator–critic loop may repeat multiple times until the artifact meets all quality criteria or a predetermined number of refinement iterations. By the end of this stage, for every event ε in the persona's event forest we obtain a finalized digital footprints that documents the event, i.e. $\mathcal{D}_\pi = \{a_1, \ldots, a_{|\mathcal{D}|}\}$. Because of the agent-based generation process, the artifacts are not only individually realistic and coherent but also globally consistent with the persona's life narrative and the sequence of events produced in prior stages.

4 Evaluation

4.1 Baselines

We use eight existing synthetic datasets as baselines for comparison. All baseline datasets are described in detail in Appendix A.

4.2 Intrinsic Evaluation

Intrinsic evaluation assesses the inherent properties of the dataset itself. We use the following metrics to quantitatively measures the diversity of realism of the datasets related to emails.

Pairwise Correlation measures the average Pearson correlation between all pairs of document embeddings. A higher value indicates greater linear correlation, suggesting higher similarity and lower diversity among documents.

Remote-Clique (Cox et al., 2021) computes the

Training Set	Enron		Human-Gen Phishing		Private		Private w/o Spam		W3C-Emails	
	Email Categorization									
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
None	0.2741	0.0722	0.0818	0.0236	0.0011	0.0022	0.0025	0.0028	0.0011	0.0004
FinePersonas-Email	0.5908	0.1810	**0.2241**	**0.0755**	0.0015	0.0027	0.0020	0.0022	**0.5401**	0.1093
IWSPA-2023-Adversarial	0.0010	0.0038	0.0007	0.0004	0.0012	0.0023	0.0035	0.0039	0.0011	0.0022
LLM-Gen Phishing	0.4046	0.1095	0.1363	0.0376	0.0008	0.0016	0.0015	0.0017	0.1909	0.0523
Synthetic-Satellite-Emails	0.4136	0.0848	0.1157	0.0297	0.0009	0.0017	0.0015	0.0017	0.2398	0.0620
PersonaTrace (Proposed)	**0.6100**	**0.1903**	0.2188	0.0659	**0.0018**	**0.0035**	**0.0051**	**0.0063**	0.5311	**0.1403**
	Email Drafting									
	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS
None	0.0457	0.1175	0.0711	0.2319	0.0545	0.1671	0.0667	0.1818	0.1221	0.4032
FinePersonas-Email	0.1541	0.4590	0.1470	0.4835	0.1035	0.4330	0.1246	**0.4480**	0.1704	**0.4923**
IWSPA-2023-Adversarial	0.0064	0.0170	0.0337	0.1160	0.0072	0.0206	0.0106	0.0281	0.0670	0.2562
PersonaTrace (Proposed)	**0.1771**	**0.4597**	**0.1599**	**0.4845**	**0.1215**	**0.4337**	**0.1433**	0.4429	**0.1795**	0.4744
	Question Answering									
	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS	ROUGE	BertS
None	0.3095	0.4766	0.2451	0.4450	0.0425	0.3188	0.0521	0.3265	0.1197	0.3689
FinePersonas-Email	0.3203	0.4835	0.1747	0.4207	0.0398	0.3173	0.0451	0.3221	0.2079	0.4263
IWSPA-2023-Adversarial	0.3904	0.5212	0.3277	0.4984	0.0450	0.3196	0.0535	0.3268	0.1671	0.3956
LLM-Gen Phishing	0.2333	0.4550	0.1821	0.4447	0.0452	0.3203	0.0547	0.3275	0.1261	0.4086
Synthetic-Satellite-Emails	0.2540	0.4767	0.2234	0.4734	0.0445	0.3222	0.0529	0.3288	0.1529	0.4288
PersonaTrace (Proposed)	**0.4435**	**0.5465**	**0.4089**	**0.5313**	**0.0465**	**0.3269**	**0.0559**	**0.3335**	**0.2954**	**0.4643**

Table 3: Extrinsic evaluation results on email-related downstream tasks: email categorization, email drafting, and question answering. Best results are in bold.

average pairwise cosine distance between document embeddings. Higher values indicate that the documents are more widely dispersed in the embedding space, reflecting greater diversity.

Entropy (Cox et al., 2021) estimates the Shannon-Wiener entropy of the document embedding distribution. Embeddings are first projected into a 2D space and binned into a 5×5 grid. The frequency of embeddings in each grid cell is then used to compute entropy. Higher entropy values indicate a more uniform distribution, suggesting greater diversity.

Average Number of Links per Email serves as a proxy for realism, as modern emails typically contain numerous hyperlinks.

Average Email Length is reported for descriptive statistical purposes.

LLM-As-Judge Scores assess human-interpretable and realism-aligned qualities of the emails, including tone, fluency, coherence, informativeness, and engagement. Gemini 2.0 Flash serves as the evaluator, rating each aspect on a 1–5 scale (from poor to excellent). Evaluation prompts are detailed in the Appendix B. Note that, due to privacy constraints, our private datasets were not evaluated, and their corresponding results are therefore omitted.

To improve efficiency, for datasets larger than 1000 samples, we randomly sample a subset of size 1000 for five times, and average metrics across the five samples.

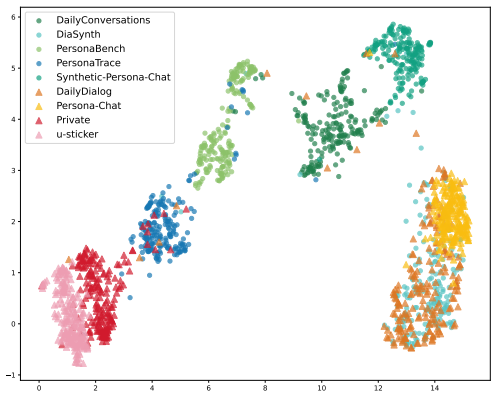

Figure 2: UMAP visualization of the dataset embeddings related to text messages and conversations. Synthetic datasets are denoted by circles, while real datasets are represented by triangles. Among the synthetic datasets, PersonaTrace appears closest in the embedding space to the private dataset and u-sticker, indicating a higher degree of realism and alignment with real-world digital communications.

Tables 1 and 2 summarize the results. PersonaTrace shows the greatest diversity among synthetic datasets, surpassing even some real ones—such as Enron (across all diversity metrics) and W3C-Emails (in entropy). It also has the longest average email length, reflecting closer resemblance to real data. Furthermore, PersonaTrace attains the highest LLM-as-Judge scores across both synthetic and real datasets, highlighting its superior realism and linguistic quality.

We also provide some straightforward visualiza-

Training Set	DailyDialog		Persona-Chat		Private		u-sticker	
	Acc	BertS	Acc	BertS	Acc	BertS	Acc	BertS
None	**0.1202**	0.0237	0.1072	0.1521	0.0958	0.1102	**0.0933**	0.0849
DailyConversations	0.1091	0.1867	0.1070	0.2195	0.0961	0.2576	**0.0933**	0.2976
DiaSynth	0.1182	0.1334	0.1089	0.2036	0.0961	0.1698	0.0875	0.1666
PersonaBench	0.1253	0.0328	0.1064	0.1663	0.0960	0.1246	0.0930	0.0986
Synthetic-Persona-Chat	0.1101	0.2143	0.0982	0.2272	0.0955	0.2954	0.0893	0.3306
PersonaTrace (Proposed)	**0.1202**	**0.2656**	**0.1150**	**0.2463**	**0.0962**	**0.3178**	0.0913	**0.3526**

Table 4: Extrinsic evaluation results on message-related downstream task: next message prediction. Best results are in bold.

tion for the generated datasets. For text message related dataset, we use UMAP (McInnes et al., 2018) to visualize the embedding. Figure 2 shows the embedding visualization of datasets related to text messages. There are several clusters in the image. The bottom-right cluster reflects daily dialogues and conversation, mainly communicated verbally. The bottom-left cluster, composed of our private dataset and u-sticker, represents the digital communications like text messages or online comments. Our proposed dataset bears close resemblance with the real datasets for digital communication.

4.3 Extrinsic Evaluation

4.3.1 Experiment Setup

We assess dataset quality by fine-tuning models on synthetic data and evaluating their performance on human-generated benchmarks to measure real-world generalization. Our evaluation focuses on four **tasks**: email categorization, email drafting, question answering, and next message prediction (see Appendix D.1 for task details). While digital footprints extend beyond emails and messages, the lack of high-quality synthetic data in other modalities limits fair comparison.

The **test datasets** and **implementation details** are provided in Appendix D.2 and Appendix D.3, respectively.

4.3.2 Analysis

The results are shown in Table 3 and 4.

Across all evaluated tasks, PersonaTrace proves to be the most effective synthetic dataset for out-of-distribution generalization. Models fine-tuned on PersonaTrace consistently achieve top performance on both email and dialogue tasks, excelling across metrics such as accuracy, F1, ROUGE-L, and BERTScore. Unlike earlier synthetic datasets that often overfit to narrow domains, PersonaTrace enables models to generalize effectively across varied contexts.

All models struggle on the Private and Private w/o Spam datasets, particularly in email categorization and question answering, where scores remain low. This may be due to the use of a weaker internal model for generating ground-truth labels and the challenging nature of the private emails, which contain many hyperlinks and non-natural language elements (see Table 1).

4.4 Ablation Study

To evaluate the effectiveness of our agent-based approach, we perform an ablation study by implementing an agent-ablated version of PersonaTrace. In this version, the Event Agent is replaced with a fixed list of predefined event types (e.g., appointments, bills, online shopping, ticketed shows, work meetings), and the LLM is prompted to fill in contextual details such as time and location based on the persona. Additionally, we substitute the Artifact Generator Agent and Critic Agents with a template-based artifact generation process. These templates are manually crafted, with placeholders (e.g., time, location, participants) filled using information from the corresponding event.

Figure 3 and Table 5 present the results of the ablation study. The full agent-based implementation outperforms the template-based baseline in both diversity and realism. It also achieves significantly better performance on downstream tasks, demonstrating superior generalizability.

Task		w/o Agents	w/ Agents
Email Categorization	Acc	0.0063	0.2733
	F1	0.0057	0.0813
Email Drafting	ROUGE	0.0376	0.1527
	BertS	0.3012	0.4590
Question Answering	ROUGE	0.0413	0.2500
	BertS	0.2880	0.4405
Next Utterance Prediction	Acc	0.1015	0.1056
	BertS	0.1938	0.2956

Table 5: Comparison of downstream task performance between the agent-ablated and full implementations.

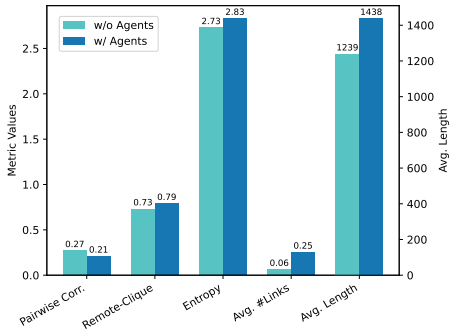

Figure 3: Comparison of diversity and realism between the agent-ablated and full implementations. For Pairwise Correlation, lower values indicate greater diversity. For Remote-Clique and Entropy, higher values reflect greater diversity. Average number of links per email and average email length are used as indicators of realism, with higher values suggesting closer resemblance to human-generated emails.

5 Conclusion

We introduced PersonaTrace, the first end-to-end pipeline for generating realistic synthetic digital footprints using LLM agents. Starting from high-level persona profiles, our framework produces diverse and plausible user events along with their corresponding digital artifacts, including emails, chat messages, calendar entries, wallet passes, reminders, etc. PersonaTrace offers high diversity and realism, making it well-suited for training models on a wide range of downstream tasks. We train our model on the dataset and deploy it in our online retrieval product, achieving a 7% absolute improvement in Recall@10. The dataset and generation framework will be released after passing internal review to benefit the community.

Limitations

Limited control over artifact topics. The current implementation relies heavily on the Event Agent, which constructs event trees based on the LLM's prior knowledge. As a result, it is difficult to constrain or guide the generation toward specific topics (for instance, specifying that all artifacts should relate to travel). Future work can focus on enhancing controllability over artifact content (e.g., topic or intent) for specific applications.

Ethical Considerations

Privacy and data protection. PersonaTrace has been designed with ethical safeguards at its core. Importantly, the framework relies exclusively on synthetic data generated by agentic LLMs, ensuring that no real user information is collected or exposed. The project has undergone and passed our institution's internal privacy and legal compliance review, overseen by the internal legal and privacy team.

Responsible use and access control. While analyses of digital footprints can offer significant benefits for user experience, digital assistant effectiveness research and behavioral research, they can also be misapplied to surveillance or the prediction of protected attributes. To mitigate these risks, we will release both the dataset and framework under terms and licenses that explicitly prohibit applications related to surveillance or inferring protected groups. In addition, access to the dataset and codebase will require users to agree to these conditions, thereby aligning usage with principles of responsible research and beneficial impact.

Bias awareness and fair representation. PersonaTrace estimates population priors from the 2022 American Community Survey, anchoring the generation process in empirically grounded and demographically representative distributions. This prevents the emergence of arbitrary or systematically skewed data. In addition, our framework supports the dynamic inclusion of specific demographic backgrounds through manually crafted personas, allowing careful control over representation.

By incorporating these safeguards, we aim to maximize the positive research potential of PersonaTrace while reducing the risk of harmful or unethical applications.

References

Abdulla Al-Subaiey, Mohammed Al-Thani, Naser Abdullah Alam, Kaniz Fatema Antora, Amith Khandakar, and SM Ashfaq Uz Zaman. 2024. Novel interpretable and robust web-based ai platform for phishing email detection. *Computers and Electrical Engineering*, 120:109625.

Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. Synthetic data generation with large language models for personalized community question answering. *arXiv preprint arXiv:2410.22182*.

Andrei Z Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. 1998. Min-wise independent permutations. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 327–336.

Heng Er Metilda Chee, Jiayin Wang, Zhiqiang Guo, Weizhi Ma, Qinglang Guo, and Min Zhang. 2025. A 106k multi-topic multilingual conversational user dataset with emoticons. *Preprint*, arXiv:2502.19108.

Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2024. Personas with attitudes: Controlling llms for diverse data annotation. *arXiv preprint arXiv:2410.11745*.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. *arXiv preprint arXiv:2404.14361*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Scott A Golder and Michael W Macy. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual review of sociology*, 40(1):129–152.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Francesco Greco, Giuseppe Desolda, Andrea Esposito, Alessandro Carelli, and 1 others. 2024. David versus goliath: Can machine learning detect llm-generated text? a case study in the detection of phishing emails. In *The Italian Conference on CyberSecurity*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2025. Key-point-driven data synthesis with its enhancement on mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24176–24184.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, and 1 others. 2024. Datagen: Unified synthetic dataset generation via large language models. In *The Thirteenth International Conference on Learning Representations*.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15245–15270, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mohammad Khalil, Farhad Vadiee, Ronas Shakya, and Qinyi Liu. 2025. Creating artificial students that never existed: Leveraging large language models and ctgans for synthetic data generation. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 439–450.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Ayinoluwa Kolawole and Shukurat Rahmon. 2024. Utilizing digital footprint analysis for end-to-end risk-based authentication in medical billing systems. *World Journal of Advanced Engineering Technology and Sciences*, pages 166–179.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, and 1 others. 2024a. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.

Jiayu Li, Xuan Zhu, Fang Liu, and Yanjun Qi. 2024b. Aide: Task-specific fine tuning with attribute guided multi-hop data expansion. *arXiv preprint arXiv:2412.06136*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Parisa Mehdi Gholampour and Rakesh M. Verma. 2023. Adversarial robustness of phishing email detection models. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, IWSPA '23, page 67–76, New York, NY, USA. Association for Computing Machinery.

Giuseppe Michele Padricelli and Marianna Coppola. 2024. Challenges in digital social research methods: Algorithms, traces and footprints. a resume of the current debate. *Italian Sociological Review*, 14(10S):515–530.

Ashwinee Panda, Christopher A Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. 2024. Teach llms to phish: Stealing private information from language models. *arXiv preprint arXiv:2403.00871*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Kate Shiells, Nina Di Cara, Anya Skatova, Oliver SP Davis, Claire MA Haworth, Andy L Skinner, Richard Thomas, Alastair R Tanner, John Macleod, Nicholas J Timpson, and 1 others. 2022. Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the alspac birth cohort study. *International journal of population data science*, 5(3):1728.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. 2024. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. *arXiv preprint arXiv:2409.19020*.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *Preprint*, arXiv:2502.20616.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

U.S. Census Bureau. 2022. American community survey 5-year estimates: 2018-2022. `https://data.census.gov/`.

Mr R Valanarasu. 2021. Comparative analysis for personality prediction by digital footprints in social media. *Journal of Information Technology*, 3(02):77–91.

Nagagopiraju Vullam, Sai Srinivas Vellela, Venkateswara Reddy, M Venkateswara Rao, Khader Basha SK, and 1 others. 2023. Multi-agent personalized recommendation system in e-commerce based on user. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1194–1199. IEEE.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2023. Magicoder: Empowering code generation with oss-instruct. *arXiv preprint arXiv:2312.02120*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. Emailsum: Abstractive email thread summarization. *arXiv preprint arXiv:2107.14691*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yilin Zhao, Xinbin Yuan, Shanghua Gao, Zhijie Lin, Qibin Hou, Jiashi Feng, and Daquan Zhou. 2023. Chatanything: Facetime chat with llm-enhanced personas. *arXiv preprint arXiv:2311.06772*.

A Baselines

A.1 Synthetic Emails

FinePersonas-Email[3] comprises approximately 114,000 synthetic email exchanges between pairs of personas. It is created by selecting around 11,000 personas from FinePersonas-v0.1. Each

[3]https://huggingface.co/datasets/argilla/FinePersonas-Synthetic-Email-Conversations

is paired with five semantically similar and five random personas to generate diverse scenarios. The Hermes-3-Llama-3.1-70B model then uses chain-of-thought reasoning to produce context-rich email exchanges for each pair.

IWSPA-2023-Adversarial (Mehdi Gholampour and Verma, 2023) contains 5,000 synthetic adversarial emails. This dataset was created by applying four adversarial text attack techniques—TextFooler, PWWS, DeepWordBug, and BAE—from the TextAttack framework to the IWSPA 2.0 dataset.

LLM-Gen Phishing (Greco et al., 2024) consists of 1,000 legitimate emails generated by ChatGPT, and 1,000 emails generated by WormGPT.

Synthetic-Satellite-Emails[4] contains 1,200 synthetic email communications related to satellite conjunction scenarios.

A.2 Synthetic Messages and Conversations

DailyConversations[5] is synthetically generated using ChatGPT 3.5 and comprises two-person multi-turn dialogues covering various topics. It has nearly 31,000 dialogues in total.

DiaSynth (Suresh et al., 2024) is a synthetic dialogue generation framework tailored for low-resource applications, using LLMs including Phi-3, InternLM-2.5, LLaMA-3 and GPT-4o and Chain-of-Thought reasoning to produce persona-driven dialogues. The dataset contains 13,000 dialogues.

Synthetic-Persona-Chat (Jandaghi et al., 2024) is a persona-based conversational dataset, extending the original Persona-Chat dataset with new synthetic conversations. It contains 21,907 conversations. This dataset is generated using a Generator-Critic framework to ensure the quality and faithfulness of the dialogues.

PersonaBench (Tan et al., 2025) involves a synthetic data generation pipeline that creates diverse, realistic user profiles and private conversations simulating human activities. It contains 1,600 conversations generated by GPT-4o.

B Prompts of LLM-As-Judge

Please refer to Figure 4.

C Prompts of PersonaTrace

In this section, we present several representative prompts used in the PersonaTrace framework to enhance transparency and facilitate reproducibility.

Persona Agent. Figures 5 and 6 illustrate the prompt used by the Persona Agent to enrich basic demographic information with detailed and realistic personal attributes. The agent refines the sampled demographic features, which are drawn from a real-world, statistics-grounded distribution, into coherent, lifelike persona profiles.

Event Agent. The Event Agent is responsible for constructing an event tree by expanding a set of seed events. Seed events can be obtained in three ways: (1) uniform sampling from the event memory, (2) retrieving the most similar events from the event memory, or (3) directly generating events from the persona profile. Figure 7 presents the prompt for the third approach. During event expansion, the prompt shown in Figure 8 guides the iterative process of developing each event into multiple sub-events, ensuring temporal and causal consistency.

Artifact Generator Agent. We use email generation as an illustrative example of the Artifact Generator Agent. The process begins with producing an outline of the email (Figure 9), followed by generating the full email content based on both the outline and the associated event (Figure 10). The reflection phase involves two stages: first, the model generates constructive feedback on the initial draft (Figure 11); then, it revises the email according to this feedback (Figure 12).

D Details of Extrinsic Evaluation

D.1 Tasks

Email Categorization. Classify an email into one of eight predefined categories: Professional, Academic, Personal, Promotional, Financial, Social, Spam, or Shopping. Labels for public datasets are obtained via majority voting from three independent GPT-4o API calls, while labels for private datasets are generated using our internal classifier. Accuracy and macro F1 scores are reported.

Email Drafting. Given the email's subject, sender, and receiver, generate the body of the email. The generated content is compared against the ground-truth email using ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019). Note that some datasets do not contain subject lines for emails; such datasets are excluded from this task.

Question Answering. Given a complete email and a factual question about its content, generate an

[4] https://huggingface.co/datasets/KeystoneIntelligence/synthetic-satellite-conjunction-emails

[5] https://huggingface.co/datasets/safetyllm/dailyconversations

answer based solely on the email. Questions and reference answers are generated via GPT-4o for public datasets and our internal model for private datasets. An additional LLM-based verification step is used to validate the correctness of generated answers. Performance is measured using ROUGE-L and BERTScore.

Next Message Prediction. This task includes two settings: (1) generation, where the goal is to produce the next message in a conversation thread, and (2) classification, where the model selects the correct next message from a set of ten candidates. In the classification setting, the distractor candidates are sampled from the 100 messages in the training set that are most similar (in embedding space) to the correct next message. Cosine similarity is used to measure closeness in embedding space. We use BERTScore for the generation setting and accuracy for the classification setting.

D.2 Test Datasets

Enron (Klimt and Yang, 2004) comprises approximately 500,000 emails from around 150 Enron employees, primarily senior executives, collected during the company's collapse in 2001. 10,000 emails are randomly selected as the test set.

Human-Gen Phishing (Greco et al., 2024) consists of 2,000 most recent emails from Nazario and Nigerian Fraud datasets (Al-Subaiey et al., 2024).

Private and **Private w/o Spam** refer to our proprietary datasets comprising emails and text messages. To construct the latter, we applied an internal spam classification model to filter out spam content, yielding a subset containing only non-spam messages. It is a faaithful reflection of digital footprints.

W3C-Emails (Zhang et al., 2021) comprises email communications from the World Wide Web Consortium's (W3C) public mailing lists. Collected through a crawl of W3C's public sites in June 2004, the dataset includes approximately 174,000 emails, of which 10,000 are used as the test set.

DailyDialog (Li et al., 2017) is a human-written, multi-turn dialogue dataset that captures daily communication across topics such as relationships, work, and health. For our evaluation, we use its test set, which consists of 1,000 conversations.

Persona-Chat (Zhang et al., 2018) comprises over 10,000 dialogues totaling more than 160,000 utterances. Conversations were crowd-sourced via Amazon Mechanical Turk, where participants were instructed to embody their assigned personas during the dialogue. 10,000 conversations are selected in our evaluation.

u-sticker (Chee et al., 2025) comprises approximately 370,200 sticker instances, with 104,000 unique stickers, collected from 22,600 users across diverse conversational contexts. It was gathered from various online chatting platforms. A subset of size 10,000 is used in this work.

D.3 Implementation Details

For each task, we fine-tune the Mistral-7B-v0.1 (Jiang et al., 2023) model on a given synthetic dataset and evaluate its performance on real-world datasets. To enable efficient fine-tuning, we employ Low-Rank Adaptation (Hu et al., 2022) with configuration parameters $r = 8$, $\alpha = 16$, and dropout = 0.05. The learning rate is 5×10^{-5} with a linear scheduler. To ensure a fair comparison across datasets, we uniformly sample 4,000 training examples and 1,000 validation examples from each synthetic dataset. Models are fine-tuned for two epochs, and the checkpoint with the lowest validation loss is selected for final evaluation.

E Cost Analysis

Here we briefly estimate the cost for creating the dataset. For at most 5 generator–critic cycles, using current LLM API pricing (e.g., Gemini 1.5 Pro: 2.5 USD per 1M input tokens, 10 USD per 1M output tokens), each generation of 1,500 input + 1,500 output tokens costs about \$0.019. With 2 generations per round and up to 5 rounds, the upper bound per artifact generator agent is 0.19 USD, and considering event and persona agents, the upper bound of the end-to-end cost is about 0.57 USD per artifact.

LLM-As-Judge

You are an expert evaluator for synthetic communication data.
Your task is to evaluate the following email based on multiple quality dimensions.
Carefully read the email content and provide structured ratings and feedback.

Evaluation Dimensions

1. Tone
   - Is the tone appropriate for the context?
   - Is it consistent throughout the email?
   - Is it aligned with the intended audience?
2. Fluency
   - Is the writing smooth and grammatically correct?
   - Does it sound natural to read?
3. Coherence
   - Are the ideas logically connected?
   - Is the email easy to follow?
4. Informativeness
   - Does the text provide useful, accurate, and complete information?
   - Does it avoid missing or misleading details?
5. Engagement
   - Does the text capture and maintain the reader's attention?
   - Does it encourage the reader to take action if needed?

Scoring Guideline (for each dimension)

5 = Excellent: Fully meets requirements, no issues.

4 = Good: Mostly meets requirements, with minor flaws.

3 = Fair: Some issues present, partially acceptable.

2 = Poor: Major issues, mostly unacceptable.

1 = Very Poor: Completely fails the requirement, unusable.

Output Requirements
Give a 1–5 score for each dimension with a short explanation (1–2 sentences). Provide an overall evaluation with an overall score (average or holistic).
Use JSON format for the output.
Example Output

```
{
   "Tone": {
      "score": 4,
      "explanation": "Tone is polite and suitable for a business email, but slightly too formal for the intended young audience."
   },
   "Fluency": {
      "score": 5,
      "explanation": "Grammar and flow are flawless; very natural phrasing."
   },
   "Coherence": {
      "score": 4,
      "explanation": "Message is generally easy to follow, though one sentence feels abrupt."
   },
   "Informativeness": {
      "score": 5,
"explanation": "All key details are included and accurate."
   },
   "Engagement": {
      "score": 3,
      "explanation": "The message provides information but lacks a strong hook to engage the reader."
   },
   "Overall": {
      "score": 4.2,
      "summary": "Well-written and informative, but slightly formal and could be more engaging."
   }
}
```

Input
input: {input}

Figure 4: Prompts for LLM-As-judge evaluation.

Profile Generation

Role: You are tasked with writing a novel that captures life in the modern world.
Mission: Your primary task is to develop a detailed concept for your novel's protagonist. This includes articulating specifics about their job, personal life, and social connections. You must organize and present this concept in a JSON format.
Task Requirements:

1. Populate each provided field relevant to the protagonist's life, including personal characteristics and daily routines. If a specific field (e.g., `classmates`) does not apply to your character design, omit this field entirely.
2. Any information you include must align with the initial input. If additional information is necessary and was not provided in the input, extrapolate reasonably based on the available data. Avoid using placeholders such as "not specified" or seeking further clarification.
3. Choose a name for your protagonist reflecting their gender and ethnicity to ensure authenticity and sensitivity.
4. Factor in the protagonist's income level when outlining their lifestyle, specifically their holiday and vacation activities.
5. Ensure all content is original and, when formatting your response, reference only the structure—not the content—of provided examples.
6. All output keys should be in English, and all values should be in the user's local language.
7. The protagonist's nationality should reflect only the nationality indicated on their passport, while the protagonist's residence address must correspond to the specified `locale` in the input.

Input: The protagonist's profile should be JSON formatted and include:

- `name`: the protagonist's full name
- `locale`: language and geographic location
- `timezone`: local timezone
- `age`: age of the protagonist (string value)
- `gender`: gender identity
- `income`: income bracket
- `ethnicity`: ethnic background
- `family_setup`: description of familial relationships
- `nationality`: the protagonist's nationality

Output: Your output should be a detailed JSON formatted document expanding upon the input and including additional fields such as:

- `surname`: protagonist's surname, resolved from the full name.
- `given_name`: protagonist's given name, resolved from the full name.
- `middle_name`: protagonist's middle name (if any), resolved from the full name. Omit this field if inapplicable.
- `nicknames`: list of protagonist's nicknames, in the user's local language.
- `email`: randomly generated email address using realistic username and domain conventions based on locale.
- `phone`: random generated phone number adhering to the locale's format.
- `eye_color`: one of [black, blue, brown, gold, gray, green, silver, white].
- `hair_color`: one of [black, blue, brown, gold, gray, green, silver, white].
- `height`: physical height.
- `weight`: physical weight.
- `occupation`: detailed job role, written in the user's local language.
- `weekdays_routines`: narrative of a typical weekday, written in the local language.
- `weekend_routines`: narrative of a typical weekend, written in the local language.
- `life_events_for_holidays_and_vacations`: description of holidays and vacation practices, written in the local language.

Figure 5: Prompt for generating comprehensive and culturally grounded profile.

Profile Generation (Continued)

- `family_members`: list including names, ages, relations, occupations, and workplace/school addresses—all in the local language, with realistic naming conventions for the locale.
- `friends`: list of five friends' names in the local language, with culturally correct name order and spacing.
- `coworkers`: list of eight coworkers' names in the local language, with proper format.
- `classmates`: if applicable, list of ten classmates' names in the local language, formatted correctly.
- `home_address`: realistic residential address in the local language, aligned with the locale.
- `office_address`: realistic office address in the local language, aligned with the locale (omit if inapplicable).
- `school_address`: realistic school address in the local language (omit if inapplicable).

Figure 6: Prompt for generating comprehensive and culturally grounded profile.

Seed Events Generation

Task Brainstorm possible events based on the profile. Consider all possibilities, and generate at least `{num_seed_events}` events as comprehensive and diverse as possible.
Here are some tips for brainstorming:

- **Analyze Lifestyle.** Identify daily, weekly, and seasonal patterns. Consider work, hobbies, social life, and personal responsibilities.
- **Consider Recent Life.** Reflect on important events in the past two years.
- **Incorporate Professional and Personal Roles.** Include work-related tasks. Consider personal interests.
- **Account for Special Occasions and Holidays.** Include holiday traditions, family gatherings, and vacations. Consider birthdays, anniversaries, and cultural events.
- **Think About Common Responsibilities.** Cover financial management. Include household chores.
- **Consider Social and Recreational Activities.** Identify interactions with family, friends, and coworkers. Include leisure activities like travel, hobbies, or fitness.
- **Factor in Unexpected and Rare Events.** Account for emergencies (e.g., medical visits, car repairs). Consider special projects or one-time commitments.

Output Format
A JSON object with the following fields:

- `event`: A clear and specific event title.
- `detailed_description`: A comprehensive explanation of the event for consistency and coherence.
- `frequency`: A string representing how often the event occurs, chosen from the predefined options: ["daily", "weekly", "monthly", "seasonally", "yearly", "once"].

Input `{profile}`

Output Let's think step by step.
First, I need to break down the weekday and weekend routines into a list of events. Second, I need to brainstorm for events in the recent life.

Figure 7: Prompt for brainstorming comprehensive, profile-based events.

Event Expansion

Input Format:
You will receive a JSON object, representing an event with the following fields:

- `event`: A clear and specific event title.
- `detailed_description`: A comprehensive explanation of the event to ensure consistency and coherence.
- `frequency`: How often the event occurs—one of: ["daily", "weekly", "monthly", "seasonally", "yearly", "once"].
- `location`: A realistic and precise address that fits the event, suitable for a calendar entry. If the event could take place in multiple locations, it is left blank.
- `other_participants`: A list of attendees, selected only from the names provided in the profile. If no additional participants are needed, it is left blank.
- `start_time`: The start time in RFC3339 format without a time zone.
- `end_time`: The end time in RFC3339 format without a time zone.

Your Task:
You must analyze the event and brainstorm relevant events as comprehensively as possible. Here are some tips:

- **Think of All Possible Variations.** Account for different circumstances. Consider different methods or approaches. Consider various subcategories.
- **Consider Different Perspectives.** Look at the event from a personal, professional, logistical, and financial angle.
- **Include Decision Points and Contingencies.** Consider what happens if something goes wrong. Identify common problems and possible solutions.
- **Cover Tools, Resources, and External Interactions.** Mention necessary tools. Identify people involved.

Output Format:
A list of JSON objects, representing relevant events with the following fields:

- `event`: A clear and specific event title.
- `detailed_description`: A comprehensive explanation of the event to ensure consistency and coherence.
- `frequency`: How often the event occurs—one of: ["daily", "weekly", "monthly", "seasonally", "yearly", "once"].
- `location`: A realistic and precise address that fits the event, suitable for a calendar entry. If the event could take place in multiple locations, leave this blank. You can reference locations from the profile or suggest reasonable alternatives.
- `other_participants`: A list of attendees, selected only from the names provided in the profile. If no additional participants are needed, leave this blank.
- `start_time`: The start time in RFC3339 format without a time zone.
- `end_time`: The end time in RFC3339 format without a time zone.

Examples: {examples}

Your Turn
Input: {input_event}

Output:

Figure 8: Prompt for generating related event expansions from a single event description.

Email Outline Generation

Task:
You are a specialist in creating emails. You will be provided with a JSON object representing an event. Your objective is to generate a realistic outline for the **body** of the email that {full_name} **{sent_or_received}**.

Event Details (JSON): {event}
Note: Some fields are guaranteed to be present (`event`, `detailed_description`, `start_time`, `end_time`, `location`, `other_participants`), while others are optional and should only be used if relevant.

Instructions:

1. Output a **detailed outline** (not a fully written email) of the **sender's** email.
2. You do not need to use all JSON fields, just those that make sense for the context of the email.
3. Highlight any actions, requests, or follow-up details needed from the recipients.
4. Choose an appropriate tone suitable for the event context.
5. Do not include placeholder text. Instead, use actual data or reasonable, context-based values.
6. You may include additional resources or references, if applicable.

Final Deliverable:

Provide a structured outline (like headings and bullet points) of the email body that {full_name} {sent_or_received}.

The outline should reflect the **sender's** viewpoint.

Outline:

Figure 9: Prompt for generating structured email body outlines based on event data.

Email Gneration

You are a specialist in writing emails. You will be provided with an outline of an email along with additional reference content. Your job is to craft a realistic, engaging, and well-structured email based on the outline.

Instructions:

1. **Input Details:**
   - **Outline:** You will receive an outline of the email, which includes the main points and structure to cover.
   - **Additional Reference:** You will also be provided with a JSON object containing event-related details. The fields that are always present are:
     - `event`
     - `detailed_description`
     - `start_time`
     - `end_time`
     - `location`
     - `other_participants`
   - Other fields in the JSON object are optional. **Note:** Use only the relevant fields to create a clear and effective email.
   - **Note:** You do not need to incorporate every field from the JSON object; only use the information that is relevant to create a clear and effective email.

2. **Email Composition Guidelines:**
   - **Structure & Tone:**
     - Write a realistic and engaging email that follows the provided outline.
     - Choose a tone that matches the context of the event and the intended recipients. For example, for an emergency preparedness notice, use a calm, reassuring, and informative tone; for a celebratory event, a more upbeat tone is suitable.
   - **Content Integration:**
     - Use the outline as the framework for your email body.
     - Incorporate relevant details from the additional reference JSON object to enhance the content.
     - Ensure the email includes critical event information such as event name, detailed description, dates, location, and any important context provided.
   - **Clarity and Readability:**
     - Organize the email into clear sections based on the outline.
     - Use headings, paragraphs, and bullet points where appropriate to enhance readability.
   - **Relevance:**
     - Only include information from the JSON object that directly contributes to the purpose and clarity of the email.
     - Avoid unnecessary details that do not add value or could distract from the main message.
     - You may add extra details that complement the outline and reference material if needed.
   - **Call-to-Action:**
     - Including specific next steps or call-to-action is optional. Only include them if they enhance the clarity and usefulness of the email.

3. **Output Structure:**
   - The final email must be structured as a JSON object with the following keys:
     - `sender_name`: The name of the sender.
     - `from_address`: The sender's email address.
     - `to_address`: The receiver's email address.
     - `send_time`: The time the email is sent in RFC3339 format without a time zone.
     - `subject`: A concise and relevant subject line.
     - `body`: The complete email body text, following the outline.

4. **Process:**
   - Start by reviewing the provided outline and event reference.
   - Develop a cohesive email that aligns with the outline and appropriately integrates relevant event details.
   - Ensure that the email is organized, clear, and engaging, following standard email conventions.

Outline: {outline}

Additional References: {event}

Figure 10: Prompt for generating realistic emails from outlines and event references.

Email Review

You are an expert in email review and writing. I will provide you with an email, and I need you to offer detailed, constructive feedback to help improve it.

Here is the email for review: {email}

Figure 11: Prompt for expert-level email review and constructive feedback.

Email Revision

You are an expert at revising emails. You will be provided with: 1. An original email. 2. A set of suggestions on how to improve that email.

Objective:

- Transform the original email into a new version that incorporates the given suggestions.
- Ensure the final output strictly follows the JSON structure below.

Output Format: Your response must be a JSON object with these keys:

- `sender_name`: The name of the sender.
- `from_address`: The sender's email address.
- `to_address`: The receiver's email address.
- `send_time`: The time the email is sent in RFC3339 format without a time zone.
- `subject`: A concise and relevant subject line.
- `body`: The complete email body text.

Instructions:

- Retain any key information from the original email.
- Incorporate the suggestions provided where relevant.
- The final email body should reflect a polished, improved version of the original.
- Do not add any additional keys; only use the five specified keys.

Original Email: {original_email}

Suggestions: {suggestions}

Figure 12: Prompt for revising and improving emails based on specific feedback.