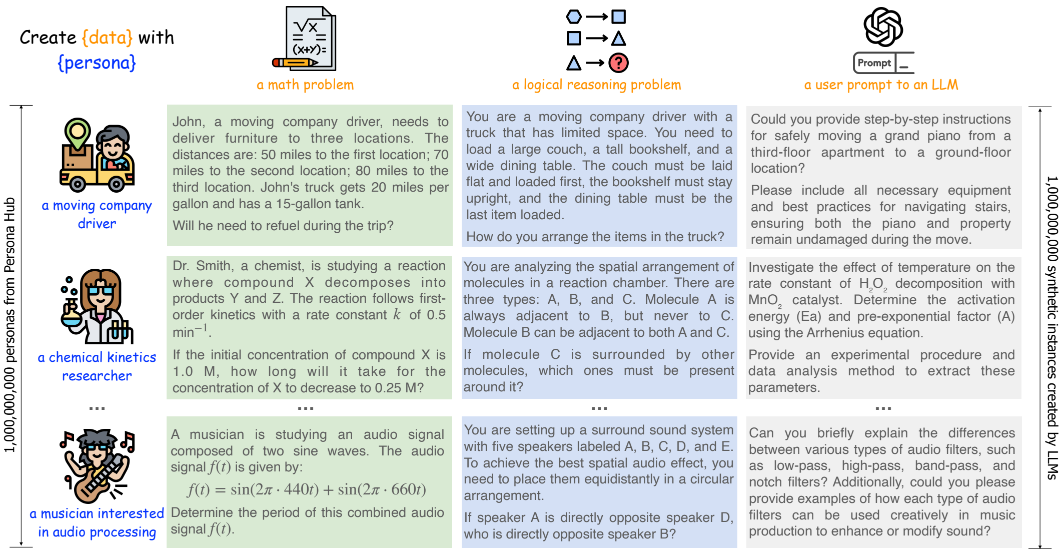
Scaling Synthetic Data Creation with 1,000,000,000 Personas

Tao Ge[*] **Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, Dong Yu**
Tencent AI Lab Seattle
https://github.com/tencent-ailab/persona-hub

Abstract

We propose a novel persona-driven data synthesis methodology that leverages various perspectives within a large language model (LLM) to create diverse synthetic data. To fully exploit this methodology at scale, we introduce Persona Hub – a collection of 1 billion diverse personas automatically curated from web data. These 1 billion personas (∼13% of the world's total population), acting as distributed carriers of world knowledge, can tap into almost every perspective encapsulated within the LLM, thereby facilitating the creation of diverse synthetic data at scale for various scenarios. By showcasing Persona Hub's use cases in synthesizing high-quality mathematical and logical reasoning problems, instructions (i.e., user prompts), knowledge-rich texts, game NPCs and tools (functions) at scale, we demonstrate persona-driven data synthesis is versatile, scalable, flexible, and easy to use, potentially driving a paradigm shift in synthetic data creation and applications in practice, which may have a profound impact on LLM research and development.

DISCLAIMER: Persona Hub *can facilitate synthetic data creation at a billion-scale to simulate diverse inputs (i.e., use cases) from a wide variety of real-world users. If this data is used as input to query a target LLM to obtain its outputs at scale, there is* ***a high risk*** *that the LLM's knowledge, intelligence and capabilities will be dumped and easily replicated, thereby challenging the leading position of the most powerful LLMs (e.g., our approach allows a 7B LLM to achieve 65% on MATH, matching the performance of* `gpt-4-turbo-preview`*). This tech report is* ***for research purposes only****. It is crucial to avoid misuse and ensure ethical and responsible application. We discuss its broad impact and potential concerns in detail in Section 5.*


Create {data} with {persona}
a math problem
a logical reasoning problem
Prompt
a user prompt to an LLM
1,000,000,000 personas from Persona Hub
1,000,000,000 synthetic instances created by LLMs

Create {data} with {persona}	a math problem	a logical reasoning problem	a user prompt to an LLM
a moving company driver	John, a moving company driver, needs to deliver furniture to three locations. The distances are: 50 miles to the first location; 70 miles to the second location; 80 miles to the third location. John's truck gets 20 miles per gallon and has a 15-gallon tank. Will he need to refuel during the trip?	You are a moving company driver with a truck that has limited space. You need to load a large couch, a tall bookshelf, and a wide dining table. The couch must be laid flat and loaded first, the bookshelf must stay upright, and the dining table must be the last item loaded. How do you arrange the items in the truck?	Could you provide step-by-step instructions for safely moving a grand piano from a third-floor apartment to a ground-floor location? Please include all necessary equipment and best practices for navigating stairs, ensuring both the piano and property remain undamaged during the move.
a chemical kinetics researcher	Dr. Smith, a chemist, is studying a reaction where compound X decomposes into products Y and Z. The reaction follows first-order kinetics with a rate constant k of 0.5 min^{-1}. If the initial concentration of compound X is 1.0 M, how long will it take for the concentration of X to decrease to 0.25 M?	You are analyzing the spatial arrangement of molecules in a reaction chamber. There are three types: A, B, and C. Molecule A is always adjacent to B, but never to C. Molecule B can be adjacent to both A and C. If molecule C is surrounded by other molecules, which ones must be present around it?	Investigate the effect of temperature on the rate constant of H_2O_2 decomposition with MnO_2 catalyst. Determine the activation energy (Ea) and pre-exponential factor (A) using the Arrhenius equation. Provide an experimental procedure and data analysis method to extract these parameters.
...	...	...	...
a musician interested in audio processing	A musician is studying an audio signal composed of two sine waves. The audio signal $f(t)$ is given by: $f(t) = \sin(2\pi \cdot 440t) + \sin(2\pi \cdot 660t)$ Determine the period of this combined audio signal $f(t)$.	You are setting up a surround sound system with five speakers labeled A, B, C, D, and E. To achieve the best spatial audio effect, you need to place them equidistantly in a circular arrangement. If speaker A is directly opposite speaker D, who is directly opposite speaker B?	Can you briefly explain the differences between various types of audio filters, such as low-pass, high-pass, band-pass, and notch filters? Additionally, could you please provide examples of how each type of audio filters can be used creatively in music production to enhance or modify sound?

Figure 1: Personas can work with a wide range of data synthesis prompts (e.g., "create a math problem") to guide an LLM to synthesize data with corresponding perspectives. The 1 billion personas in Persona Hub can facilitate various data synthesis scenarios at a billion scale.

---

[*] Project lead

1 Introduction

As synthetic data (Bauer et al., 2024; Liu et al., 2024), typically referring to data generated by models or algorithms rather than directly by humans, becomes increasingly valued (Li et al., 2023b) for training large language models (LLMs), there is a growing interest in data synthesis using LLMs: by simply specifying a data synthesis prompt, an LLM is expected to produce desirable synthetic data.

In practice, however, it is non-trivial to create synthetic data at scale: while we can easily scale up the quantity of synthetic data, it is difficult to ensure its diversity scales up as well. Without considering sampling[1], an LLM can only produce 1 instance given a data synthesis prompt. Therefore, to create diverse synthetic data at scale (e.g., 1 billion diverse math problems), a large number of diverse prompts are needed.

Previous research tends to diversify the data synthesis prompt through the following two paradigms, but unfortunately, neither can practically achieve scalable synthetic data creation:

- **Instance-driven**: This approach diversifies the data synthesis prompt by leveraging a seed corpus (i.e., creating new instances based on the instances in the seed corpus). Representative studies include Wang et al. (2022) and Yu et al. (2023). However, under this paradigm, the diversity of the synthesized data mainly comes from the seed instances, making it difficult to truly extend beyond the seed corpus. Given the limited size of a seed corpus in most practical scenarios, it is challenging for this paradigm to scale up the creation of synthetic data.
- **Key-point-driven**: This approach diversifies the data synthesis prompt with a curated comprehensive list of key points (or concepts) that can be a topic, a subject, or any knowledge we expect synthetic data to encompass. Representative studies include Li et al. (2024b) and Huang et al. (2024). However, this methodology also faces difficulties in scaling synthetic data creation: it is practically prohibitive to curate a comprehensive list by enumerating all key points across different levels of granularity, unless limited to a narrow and specific domain (e.g., mathematics).

To practically achieve diverse synthetic data creation at scale, we propose a novel **persona-driven** data synthesis methodology. This is inspired by the observation that simply adding a persona to a data synthesis prompt can steer the LLM towards the corresponding perspective to create distinctive synthetic data, as shown in Figure 1. Since almost any LLM use case can be associated with a specific persona, we can create all-encompassing synthetic data at scale as long as we construct a comprehensive persona collection.

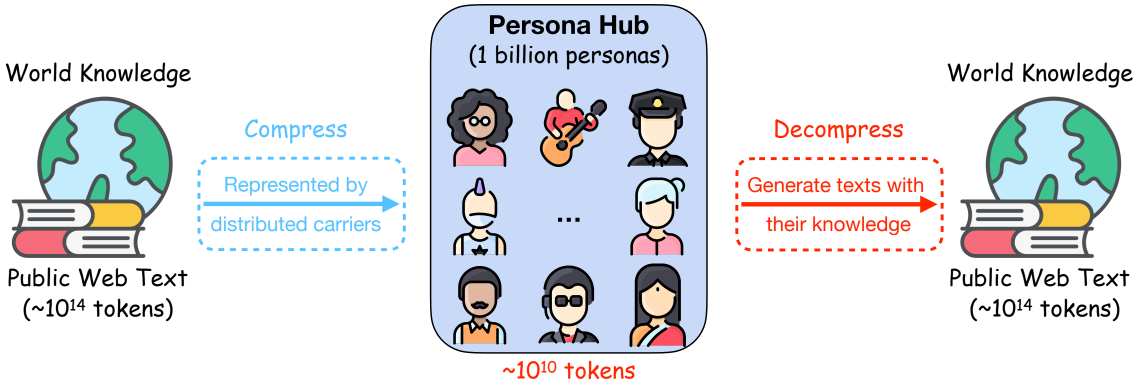


Figure 2: From a compression perspective (Delétang et al., 2023; Ge et al., 2024), Persona Hub ($\sim 10^{10}$ tokens) can be seen as the compressed form of world knowledge (public web text for training LLMs, $\sim 10^{14}$ tokens) into distributed carriers. On the other hand, the public web text can be seen as the decompressed content created by these personas with their knowledge and experiences.

---

[1] Sampling is orthogonal to this work. The diversity it introduces when solely used for data synthesis is usually limited.

Fortunately, personas are very easy to scale up. From massive web data, we automatically construct Persona Hub — a persona collection containing 1 billion diverse personas ($\sim$13% of the world's total population). As Figure 2 shows, these 1 billion personas can be regarded as distributed carriers of world knowledge, and each individual can be associated with their unique knowledge, experience, interest, personality and profession; thus, they can tap into almost every perspective encapsulated within the LLM to create diverse synthetic data at scale, without being limited by the size of a seed corpus. Moreover, in contrast to key points that typically work with specific data synthesis prompts, personas can be combined with almost any data synthesis prompt, benefiting from an LLM's strong roleplay ability (Shanahan et al., 2023; Li et al., 2023a; Choi & Li, 2024; Wang et al., 2024), making them generally applicable to a variety of data synthesis scenarios.

We showcase Persona Hub's use cases in large-scale creation of math and logical reasoning problems, instructions (i.e., user prompts), broad-coverage knowledge-rich texts, game NPCs, and tool (function) development. We demonstrate that persona-driven data synthesis is versatile, scalable, flexible, and easy to use, potentially driving a paradigm shift in synthetic data creation and applications in practice, which may have a profound impact on LLM research and development.

To facilitate research in persona-driven data synthesis, we initially release **200,000 personas** from Persona Hub and following synthetic data samples we created with various personas, including:

- **50,000 math problems**
- **50,000 instructions**
- **10,000 game NPCs**
- **50,000 logical reasoning problems**
- **10,000 knowledge-rich texts**
- **5,000 tools (functions)**

We are open to releasing more data when we can better assess the potential risks and concerns, which will be discussed in detail in Section 5.

Note: Our proposed methodology is applicable to almost any popular LLM[2]. The prompts shown in the figures throughout this paper are not exactly the prompt strings we used in our experiments; instead, they are simplified to fit the space and better illustrate the concepts. Interested readers can easily verify our methodology using the persona samples we have released. It is also worth noting that the main focus of this work is on creating new synthetic data, unlike much previous research that focuses on generating synthetic outputs for specific inputs (e.g., a math problem). Therefore, we use the terms "create" and "synthesize" interchangeably throughout the paper.

2 Persona Hub

We propose two scalable approaches to derive diverse personas to construct Persona Hub from massive web data: *Text-to-Persona* and *Persona-to-Persona*.

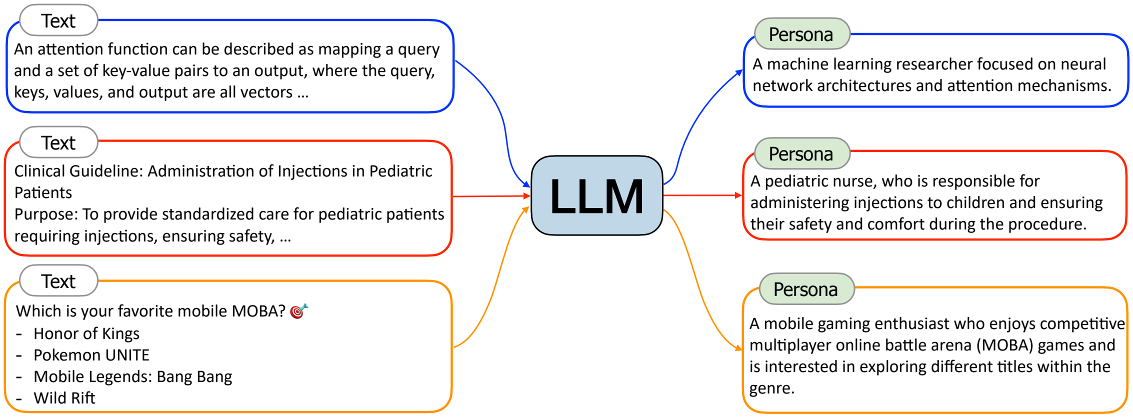
2.1 Text-to-Persona

A person with specific professional experiences and cultural backgrounds will have unique interests in reading and writing. Therefore, from a specific text, we can infer a specific persona who is likely to [`read`|`write`|`like`|`dislike`|`...`] the text. Given that text data on the web is virtually unlimited and all-encompassing, we can obtain a wide-ranging collection of personas simply by prompting an LLM with these web texts, as shown in Figure 3.

There are many formats (e.g., plain text or structured text) to represent a persona, which can be controlled within the prompt. The granularity of an output persona description can also be adjusted through the prompt. For example, in the first case, a coarse-grained persona might be "*a computer scientist*", whereas the fine-grained persona is "*a machine learning researcher focused on neural network architectures and attention mechanisms*". In our practice, we ask the LLM (in the prompt) to output

---

[2] We mainly use publicly available LLMs such as GPT-4 (Achiam et al., 2023), Llama-3 and Qwen (Team, 2024; qwe, 2024) in our experiments.



An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors …

Clinical Guideline: Administration of Injections in Pediatric Patients
Purpose: To provide standardized care for pediatric patients requiring injections, ensuring safety, …

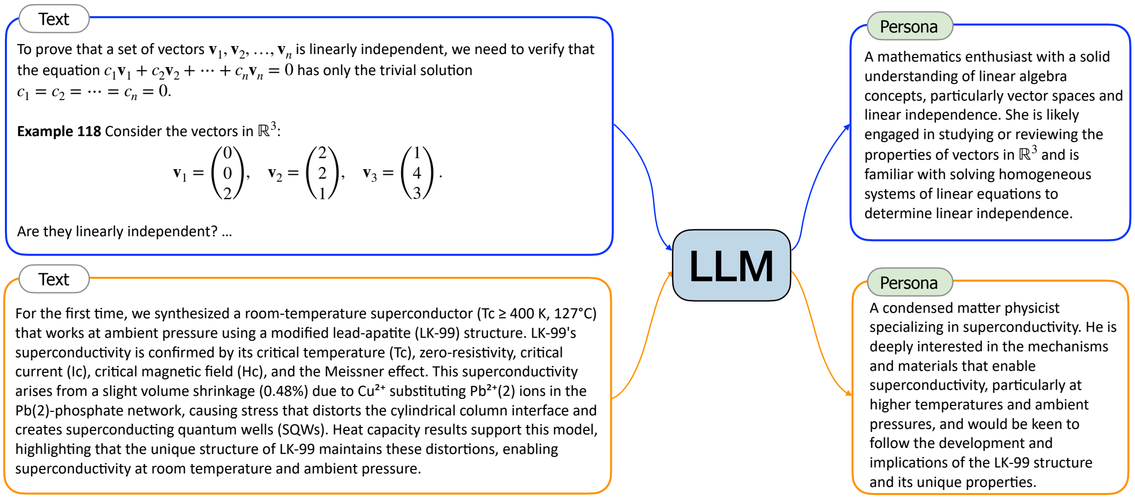
Which is your favorite mobile MOBA? 🎯
- Honor of Kings
- Pokemon UNITE
- Mobile Legends: Bang Bang
- Wild Rift

A machine learning researcher focused on neural network architectures and attention mechanisms.

A pediatric nurse, who is responsible for administering injections to children and ensuring their safety and comfort during the procedure.

A mobile gaming enthusiast who enjoys competitive multiplayer online battle arena (MOBA) games and is interested in exploring different titles within the genre.

Figure 3: The *Text-to-Persona* approach: it can use any text as input to obtain corresponding personas just by prompting the LLM "Who is likely to [`read`|`write`|`like`|`dislike`|`...`] the text?"



To prove that a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is linearly independent, we need to verify that the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = 0$ has only the trivial solution $c_1 = c_2 = \cdots = c_n = 0$.

Example 118 Consider the vectors in $\mathbb{R}^3$:

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix}.$$

Are they linearly independent? …

For the first time, we synthesized a room-temperature superconductor (Tc ≥ 400 K, 127°C) that works at ambient pressure using a modified lead-apatite (LK-99) structure. LK-99's superconductivity is confirmed by its critical temperature (Tc), zero-resistivity, critical current (Ic), critical magnetic field (Hc), and the Meissner effect. This superconductivity arises from a slight volume shrinkage (0.48%) due to Cu^{2+} substituting Pb^{2+}(2) ions in the Pb(2)-phosphate network, causing stress that distorts the cylindrical column interface and creates superconducting quantum wells (SQWs). Heat capacity results support this model, highlighting that the unique structure of LK-99 maintains these distortions, enabling superconductivity at room temperature and ambient pressure.

A mathematics enthusiast with a solid understanding of linear algebra concepts, particularly vector spaces and linear independence. She is likely engaged in studying or reviewing the properties of vectors in $\mathbb{R}^3$ and is familiar with solving homogeneous systems of linear equations to determine linear independence.

A condensed matter physicist specializing in superconductivity. He is deeply interested in the mechanisms and materials that enable superconductivity, particularly at higher temperatures and ambient pressures, and would be keen to follow the development and implications of the LK-99 structure and its unique properties.

Figure 4: Persona descriptions will be fine-grained if input texts involve many detailed elements.

persona descriptions as specifically as possible. Besides specifying the granularity of persona descriptions in the prompt, input texts can also influence the granularity of persona descriptions. As shown in Figure 4, if an input text (e.g., from a mathematical textbook or an academic paper about superconductivity) contains many detailed elements, the resulting persona description will also be specific and fine-grained. Therefore, by applying the *Text-to-Persona* approach to massive web text data, we can obtain billions (or even trillions) of diverse personas, encompassing a wide range of aspects across different granularities.

2.2 Persona-to-Persona

As discussed above, *Text-to-Persona* is a highly scalable method that can synthesize personas covering almost every aspect. However, it may still miss some personas that have low visibility on the web and thus are less likely to obtain via *Text-to-Persona*, such as *a child*, *a beggar*, or *a behind-the-scenes crew member of a movie*. To supplement the personas that *Text-to-Persona* might hardly reach, we propose *Persona-to-Persona*, which derives personas with interpersonal relationships from those obtained through *Text-to-Persona*.

As shown in Figure 5, the persona about "*a child*" can be derived from the persona of a nurse at a children's hospital (patient-caregiver relationship). Similarly, "*a beggar*" can be derived from the

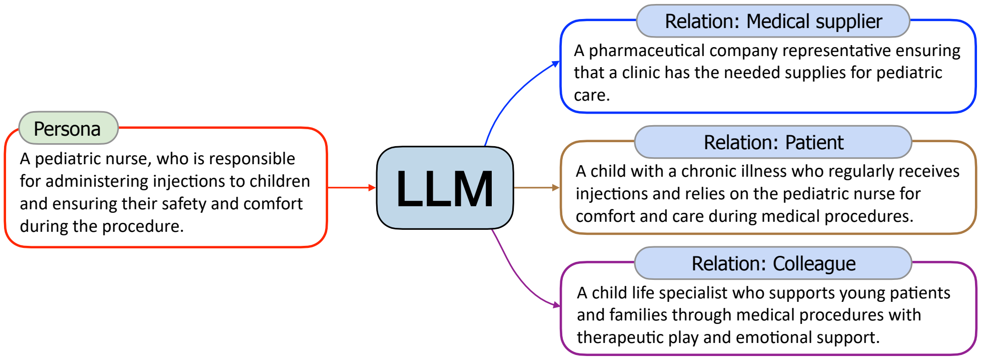

Persona
A pediatric nurse, who is responsible for administering injections to children and ensuring their safety and comfort during the procedure.
LLM
Relation: Medical supplier
A pharmaceutical company representative ensuring that a clinic has the needed supplies for pediatric care.
Relation: Patient
A child with a chronic illness who regularly receives injections and relies on the pediatric nurse for comfort and care during medical procedures.
Relation: Colleague
A child life specialist who supports young patients and families through medical procedures with therapeutic play and emotional support.

Figure 5: *Persona-to-Persona* obtains diverse personas via interpersonal relationships, which can be easily achieved by prompting the LLM "Who is in close relationship with the given persona?"

persona of a shelter worker (assistance relationship), and "*a behind-the-scenes movie crew member*" can be derived from the persona of the movie's lead actor (co-worker relationship). According to the six degrees of separation theory (Travers & Milgram, 1977), we perform six iterations of persona relationship expansion for each persona obtained through *Text-to-Persona*, thereby enriching our persona collection even further.

2.3 Deduplication

We first run *Text-to-Persona* on the RedPajama v2 dataset (Computer, 2023) and then perform *Persona-to-Persona*, as described in Sections 2.1 and 2.2. After obtaining billions of personas, it is inevitable that some of the personas will be identical or extremely similar. To ensure the diversity of Persona Hub, we deduplicate these personas in two ways:

MinHash-based Deduplication We use MinHash (Broder, 1997) to deduplicate based on the n-gram features of persona descriptions. Since persona descriptions are usually just 1-2 sentences, much shorter than a document, we simply used 1-gram and a signature size of 128 for MinHash deduplication. We deduplicate at the similarity threshold of 0.9.

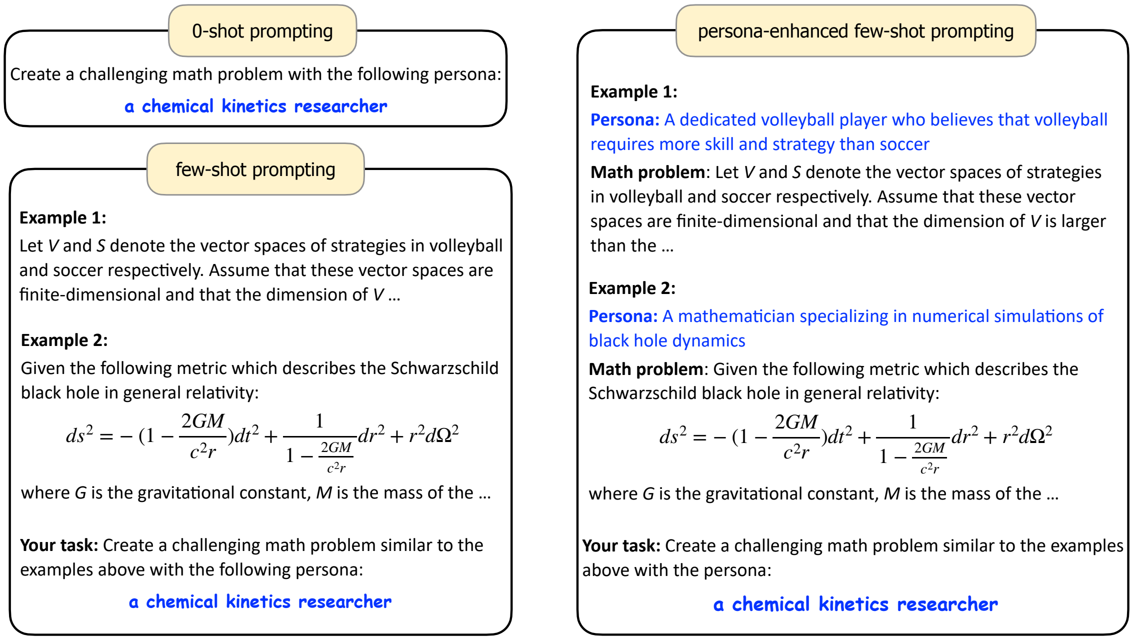
Embedding-based Deduplication After deduplication based on surface forms (i.e., MinHash with n-gram features), we also adopt embedding-based deduplication. We use a text embedding model (e.g., the `text-embedding-3-small` model from OpenAI) to compute an embedding for each persona, and then filter out personas with a cosine semantic similarity greater than 0.9.

Note that although we select 0.9 as the threshold here, we can flexibly adjust it according to specific needs for further deduplication. For instance, when the requirement for the number of instances is not high (e.g., only needing 1 million instances) but the demand for diversity is high, we can further apply a stricter deduplication standard (e.g., discarding personas with a similarity greater than 0.5).

After deduplication and using simple heuristic methods to filter out low-quality persona descriptions, we have harvested a total of 1,015,863,523 personas, finally forming our Persona Hub.

3 Persona-driven Synthetic Data Creation

Our proposed persona-driven data synthesis approach is straightforward and effective, which involves integrating a persona into the appropriate position in a data synthesis prompt. Simple as it appears, it can significantly influence the LLM to adopt the persona's perspective to create synthetic data. Driven by the 1 billion personas in Persona Hub, this approach can easily create diverse synthetic data at a billion scale.



0-shot prompting

Create a challenging math problem with the following persona:

a chemical kinetics researcher

few-shot prompting

Example 1:

Let V and S denote the vector spaces of strategies in volleyball and soccer respectively. Assume that these vector spaces are finite-dimensional and that the dimension of V ...

Example 2:

Given the following metric which describes the Schwarzschild black hole in general relativity:

$$ds^2 = -(1 - \frac{2GM}{c^2 r})dt^2 + \frac{1}{1 - \frac{2GM}{c^2 r}}dr^2 + r^2 d\Omega^2$$

where G is the gravitational constant, M is the mass of the ...

Your task: Create a challenging math problem similar to the examples above with the following persona:

a chemical kinetics researcher

persona-enhanced few-shot prompting

Example 1:

Persona: A dedicated volleyball player who believes that volleyball requires more skill and strategy than soccer

Math problem: Let V and S denote the vector spaces of strategies in volleyball and soccer respectively. Assume that these vector spaces are finite-dimensional and that the dimension of V is larger than the ...

Example 2:

Persona: A mathematician specializing in numerical simulations of black hole dynamics

Math problem: Given the following metric which describes the Schwarzschild black hole in general relativity:

$$ds^2 = -(1 - \frac{2GM}{c^2 r})dt^2 + \frac{1}{1 - \frac{2GM}{c^2 r}}dr^2 + r^2 d\Omega^2$$

where G is the gravitational constant, M is the mass of the ...

Your task: Create a challenging math problem similar to the examples above with the persona:

a chemical kinetics researcher

Figure 6: 0-shot, few-shot and persona-enhanced few-shot prompting methods.

Just as we can use either zero-shot or few-shot methods to prompt an LLM, the persona-driven methodology is also flexible and compatible with various forms of prompts to create synthetic data. As shown in Figure 6, we propose three persona-driven data synthesis prompting methods:

- **Zero-shot prompting** does not leverage any existing examples (i.e., demonstrations), thereby fully exploiting the model's creativity without being constrained by specific examples.
- **Few-shot prompting** can better ensure that the synthesized data meets the requirements by providing some demonstrations.
- **Persona-enhanced few-shot prompting** is more effective in enhancing the LLM's persona-driven data synthesis capabilities. However, its drawback is that it requires deriving the corresponding persona for each demonstration in the few-shot prompt beforehand.

4 Use Cases

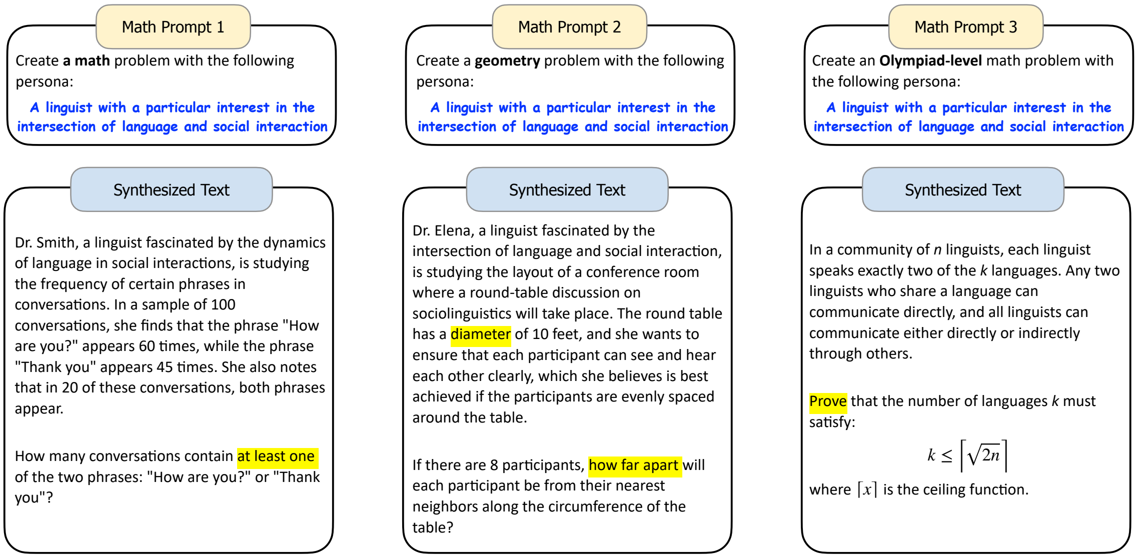
We demonstrate the use cases of Persona Hub in various data synthesis scenarios, including the large-scale creation of math and logical reasoning problems, instructions (i.e., user prompts), knowledge-rich texts, game NPCs, and tool (function) development.

As mentioned earlier, the persona-driven approach is general and versatile, making it easily adaptable to different data synthesis scenarios simply by adjusting the data synthesis prompt. Therefore, we will provide a detailed technical discussion only for math problem synthesis (Section 4.1) and skip the detailed discussion for other use cases.

4.1 Math Problems

4.1.1 Demonstrations

As the initial example (Figure 1) shows, when prompting an LLM to create a math problem, adding a persona leads the LLM to create math problems related to that persona. The example in Figure 7(left) further confirms this: when presented with a linguist persona, the LLM will create a math problem in the context of computational linguistics. Moreover, adding a persona does not hinder



Math Prompt 1

Create a **math** problem with the following persona:

A linguist with a particular interest in the intersection of language and social interaction

Synthesized Text

Dr. Smith, a linguist fascinated by the dynamics of language in social interactions, is studying the frequency of certain phrases in conversations. In a sample of 100 conversations, she finds that the phrase "How are you?" appears 60 times, while the phrase "Thank you" appears 45 times. She also notes that in 20 of these conversations, both phrases appear.

How many conversations contain at least one of the two phrases: "How are you?" or "Thank you"?

Math Prompt 2

Create a **geometry** problem with the following persona:

A linguist with a particular interest in the intersection of language and social interaction

Synthesized Text

Dr. Elena, a linguist fascinated by the intersection of language and social interaction, is studying the layout of a conference room where a round-table discussion on sociolinguistics will take place. The round table has a diameter of 10 feet, and she wants to ensure that each participant can see and hear each other clearly, which she believes is best achieved if the participants are evenly spaced around the table.

If there are 8 participants, how far apart will each participant be from their nearest neighbors along the circumference of the table?

Math Prompt 3

Create an **Olympiad-level** math problem with the following persona:

A linguist with a particular interest in the intersection of language and social interaction

Synthesized Text

In a community of n linguists, each linguist speaks exactly two of the k languages. Any two linguists who share a language can communicate directly, and all linguists can communicate either directly or indirectly through others.

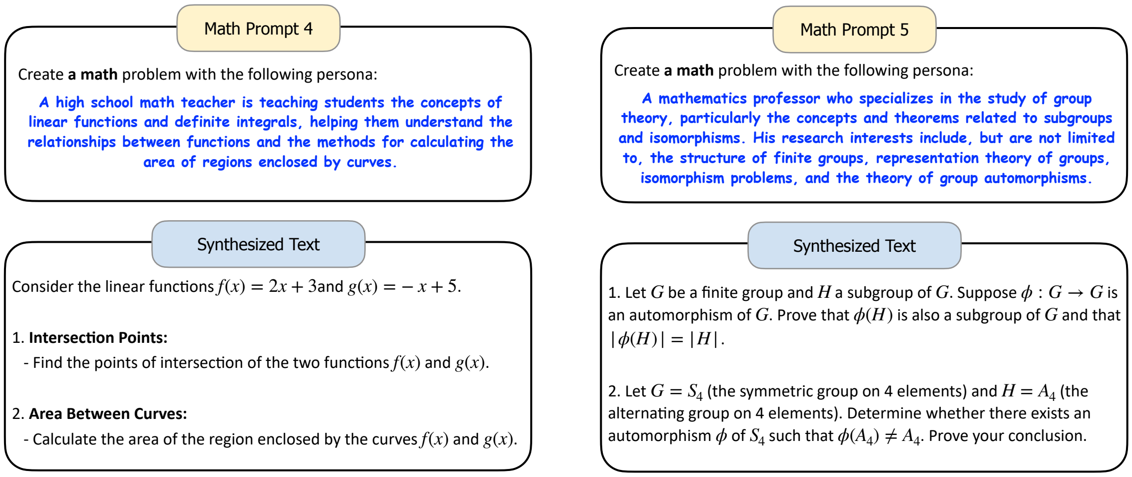
Prove that the number of languages k must satisfy:

$$k \leq \left\lceil \sqrt{2n} \right\rceil$$

where $\lceil x \rceil$ is the ceiling function.

Figure 7: A linguist persona with different math problem creation prompts that specify the focus (e.g., geometry) or the difficulty (e.g., Olympiad-level)

the flexibility of the prompt – we can still easily specify the focus (Figure 7(middle)) or difficulty (Figure 7(right)) of our desired math problem in the prompt.



Math Prompt 4

Create **a math** problem with the following persona:

A high school math teacher is teaching students the concepts of linear functions and definite integrals, helping them understand the relationships between functions and the methods for calculating the area of regions enclosed by curves.

Synthesized Text

Consider the linear functions $f(x) = 2x + 3$and $g(x) = -x + 5$.

1. **Intersection Points:**
   - Find the points of intersection of the two functions $f(x)$ and $g(x)$.

2. **Area Between Curves:**
   - Calculate the area of the region enclosed by the curves $f(x)$ and $g(x)$.

Math Prompt 5

Create **a math** problem with the following persona:

A mathematics professor who specializes in the study of group theory, particularly the concepts and theorems related to subgroups and isomorphisms. His research interests include, but are not limited to, the structure of finite groups, representation theory of groups, isomorphism problems, and the theory of group automorphisms.

Synthesized Text

1. Let G be a finite group and H a subgroup of G. Suppose $\phi : G \to G$ is an automorphism of G. Prove that $\phi(H)$ is also a subgroup of G and that $|\phi(H)| = |H|$.

2. Let $G = S_4$ (the symmetric group on 4 elements) and $H = A_4$ (the alternating group on 4 elements). Determine whether there exists an automorphism ϕ of S_4 such that $\phi(A_4) \neq A_4$. Prove your conclusion.

Figure 8: Examples of math problems created with personas of professionals related to the field of mathematics. They tend to be more challenging than those created with general personas because they usually require a deeper and more fine-grained understanding of advanced mathematical knowledge and skills.

The examples in Figure 7 demonstrate the use of general personas to create math problems. We can certainly employ professionals related to mathematics to create math problems as well. As shown in Figure 8, personas of math professionals[3] often mention more advanced and granular mathematics knowledge and skills (as discussed earlier in Section 2.1 and Figure 4), which in turn allows the created math problems to cover these mathematical concepts, making them more challenging.

---

[3] We can easily harvest a large number of such personas when running *Text-to-Persona* (Section 2.1) on public web texts, particularly when processing texts in the field of mathematics.

4.1.2 Evaluation

Data We select[4] 1.09 million personas from Persona Hub and employ the 0-shot prompting method using GPT-4 to create math problems with these personas, which does not leverage any instances from benchmarks like MATH (Hendrycks et al., 2021) during the creation of math problems. This approach allowed us to synthesize 1.09M math problems. Since this work focuses on creating new synthetic data rather than synthesizing solutions, we simply used `gpt-4o` (assistant[5]) to generate solutions to the created problems. Among these 1.09M math problems, we randomly hold out 20k as a synthetic test set to facilitate evaluation. The remaining 1.07M problems are used for training.

Test sets We use the following two test sets for evaluation:

- **Synthetic Test Set** (In-distribution): Since the set of the held-out 20K problems is produced in the same way as the 1.07M training instances, it can be considered an in-distribution test set. To ensure the accuracy of the answers in this test set for increasing the reliability of the evaluation, we additionally generate solutions using `gpt-4o` (PoT[6]) and `gpt-4-turbo` (assistant) in addition to the solution generated by `gpt-4o` (assistant). We retain only the test instances where at least two solutions are consistent. The remaining test set consists of 11.6K test instances.
- **MATH** (Out-of-distribution): The most widely recognized benchmark for testing the mathematical reasoning ability of LLMs. Its test set contains 5,000 competitive-level math problems with reference answers. Since we do not use any instances from the MATH dataset for data synthesis or training, we regard the MATH test set as an out-of-distribution test set.

Equality Checking We follow the same evaluation protocol as OpenAI[7] to check answer equality on the MATH benchmark. For the synthetic test set, we use a similar method, except we use Llama-3-70B-Instruct instead of `gpt-4-turbo-preview` as the equality checker.

We simply fine-tune the latest open-sourced 7B LLM – Qwen2-7B (qwe, 2024) with our synthesized 1.07 million math problems and evaluate its greedy decoding outputs on the above two test sets.

Model	**Model Size**	**Accuracy (%)**
Open-sourced LLMs		
DeepSeek LLM 67B Chat (Bi et al., 2024)	67B	53.2
Phi-3-Mini-4K-Instruct (Abdin et al., 2024)	3.8B	68.3
Yi-1.5-34B-Chat (Young et al., 2024)	34B	70.4
Qwen1.5-72B-Chat (Team, 2024)	72B	60.7
Qwen1.5-110B-Chat (Team, 2024)	110B	73.0
Qwen2-7B-Instruct (qwe, 2024)	7B	72.1
Qwen2-72B-Instruct (qwe, 2024)	72B	77.2
Llama-3-8B-Instruct	8B	39.8
Llama-3-70B-Instruct	70B	63.5
GPT-4		
`gpt-4-turbo-2024-04-09`	?	88.1
`gpt-4o-2024-05-13`	?	91.2
This work		
Qwen2-7B (fine-tuned *w/* the 1.07M synthesized instances)	**7B**	**79.4**

Table 1: In-distribution evaluation results on the 11.6K synthetic test instances.

---

[4] Technically, if we use the full version of Persona Hub, we can obtain 1 billion math problems synthesized by an LLM. However, due to the cost of GPT-4 APIs, we limit our scaling to 1.09M personas for this experiment.

[5] Assistant system message in OpenAI API doc: "You are a helpful assistant."

[6] Program of thought prompting (Chen et al., 2022)

[7] `https://github.com/openai/simple-evals`

Model	**Model Size**	**Accuracy (%)**
State-of-the-art LLMs		
`gpt-4o-2024-05-13`	?	**76.6**
`gpt-4-turbo-2024-04-09`	?	73.4
`gpt-4-turbo-0125-preview`	?	64.5
`gpt-4-turbo-1106-preview`	?	64.3
`gpt-4`	?	52.6*
Claude 3.5 Sonnet	?	71.1*
Claude 3 Opus	?	63.8
Gemini Pro 1.5 (May 2024)	?	67.7*
Gemini Ultra	?	53.2*
DeepSeek-Coder-V2-Instruct (Zhu et al., 2024)	236B/21B	75.7*
Llama-3-70B-Instruct	70B	52.8
Qwen2-72B-Instruct	72B	59.7*
Qwen2-7B-Instruct	7B	49.6*
This work		
Qwen2-7B (fine-tuned *w/* the 1.07M synthesized instances)	**7B**	**64.9**

Table 2: Out-of-distribution evaluation on MATH. Results marked with an asterisk (*) may not use the OpenAI's evaluation method. The model fine-tuned with our synthesized 1.07M math problems achieves 64.9% on MATH, matching the performance of `gpt-4-turbo-preview` at only a 7B scale.

Table 1 presents the in-distribution (ID) evaluation results on the 11.6K synthetic test instances. Among the tested open-source LLMs, Qwen2-72B-Instruct achieves the best result, and the ranking of the other models is generally consistent with their reported performance on other mathematical benchmarks. Our model, with the help of the 1.07M synthetic math problems, achieves nearly 80% accuracy, surpassing all the open-source LLMs. However, considering that the answers in the synthetic test are not absolutely reliable and that our model might be the only one using ID training data, this ID evaluation results should be taken as a reference only.

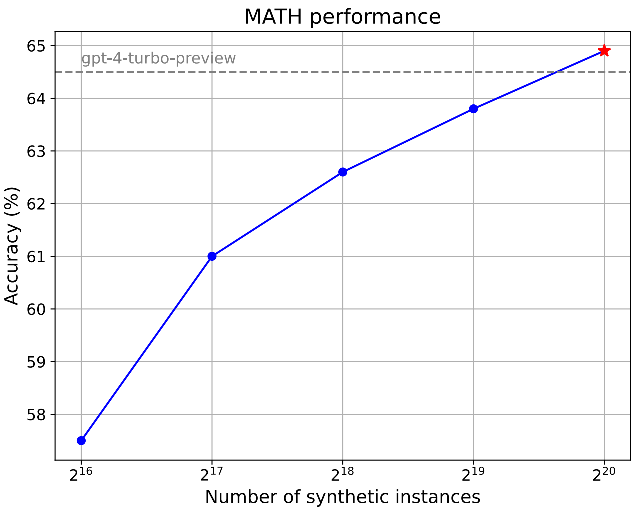
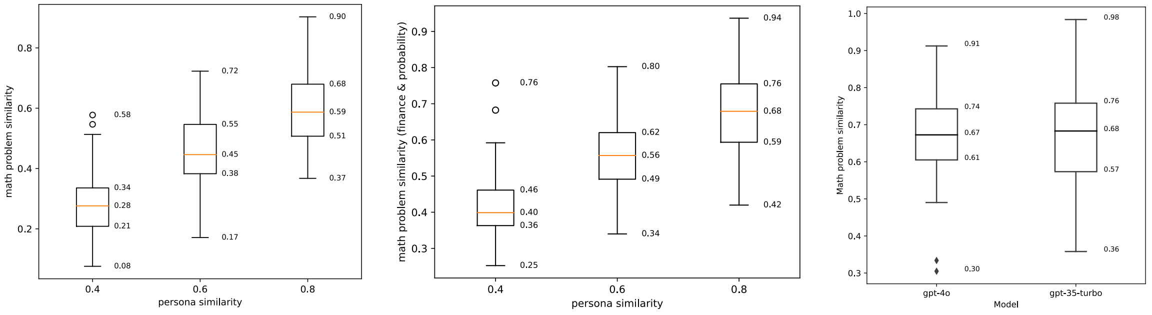


Figure 9: Accuracy on MATH with scaling the synthetic instances used for training Qwen2-7B

We present the evaluation results on MATH in Table 2. The 7B model fine-tuned with the synthetic training data achieved an impressive 64.9% accuracy on MATH simply using greedy decoding, outperformed only by `gpt-4o`, `gpt-4-turbo-2024-04-09`, Claude 3.5 Sonnet, Gemini Pro 1.5 (May 2024) and DeepSeek-Coder-V2-Instruct.

Figure 9 presents the performance of the model on MATH when trained with synthetic math problems at different scales. Its performance trend generally aligns with the scaling law (Kaplan et al., 2020). Unlike previous research (Yu et al., 2023; Wang et al., 2023; Li et al., 2024a) that performs scaling on in-distribution data (e.g., heavily relying on MATH train data to augment in-distribution data), we did not use any instances from MATH during data synthesis or training. Therefore, in this out-of-distribution (OOD) evaluation setting, achieving performance on MATH that surpasses `gpt-4-turbo-preview (1106/0125)` is indeed impressive and promising for a 7B model.

We examine the quality of our synthesized math problems: we sample 200 challenging problems (involving high school and university-level math knowledge points in China), and have two math



(a) (b) (c)

Figure 10: Similarities of math problems created by personas with different similarities: **(a)** Similarity of math problems when no specific focus is given; **(b)** Similarity of math problems when the prompt specifies they must be related to finance and probability; **(c)** Similarity of math problems synthesized by `gpt-4o` and `gpt-35-turbo` with persona similarity of 0.9.

experts evaluate their validity. Only 7 out of 200 problems are marked as invalid (e.g., due to insufficient or conflicting conditions), yielding a reliable validity rate of 96.5%.

Moreover, we specifically examine the impact of differences in personas within the prompts on the synthesized math problems. We first sample 100 pairs of personas with semantic similarities[8] of 0.4, 0.6, and 0.8, respectively. For each pair of personas, we use them to create a pair of math problems using greedy decoding (i.e., temperature=0). Then, we compute the semantic similarity of these math problem pairs and show the results in Figure 10.

We can clearly observe that the semantic similarity between synthesized math problems tends to be correlated with but lower than the similarity between their corresponding personas. When we add more specific constraints to the prompts (e.g., math problems about finance and probability), the similarity between the synthesized math problems tends to become higher (Figure 10(b)). In Figure 10(c), we also test the similarity of math problems created by `gpt-4o` and `gpt-35-turbo` using highly similar personas (similarity=0.9). The results indicate that the semantic similarity of the math problems created by `gpt-4o` and `gpt-35-turbo` seems not significantly different: most synthesized math problems' similarity falls within the range of 0.6 to 0.75, which is much lower than the similarity of the personas (0.9). Given these observations, we believe that using the personas in Persona Hub can ensure the diversity of synthesized data – even at a billion scale.

4.2 Logical Reasoning Problems

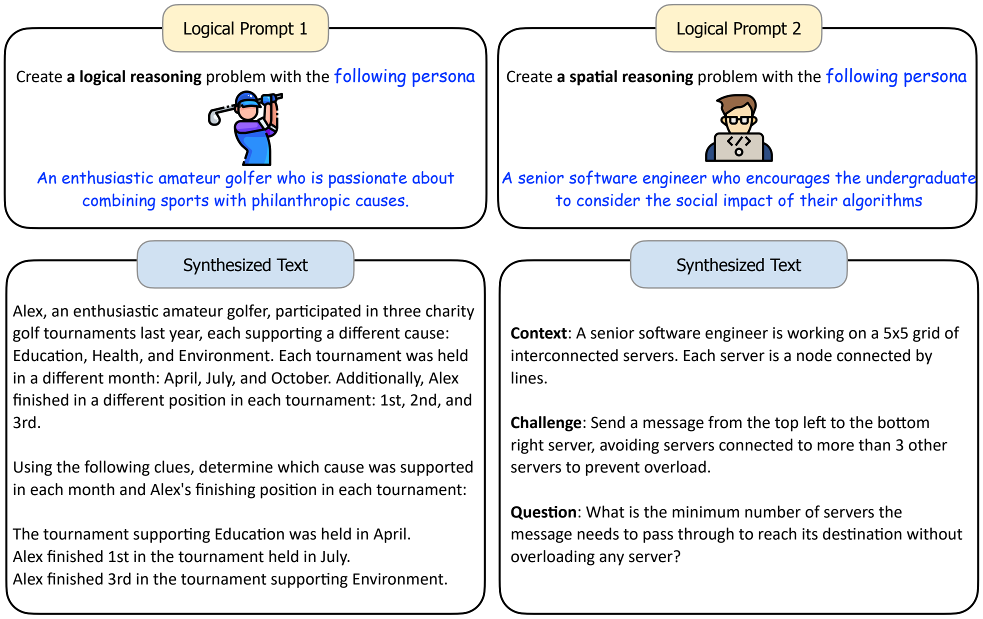
Similar to math problems, logical reasoning problems can also be easily synthesized. We present examples of typical logical reasoning problems synthesized using our proposed persona-driven methodology in Figure 11.

Moreover, we also show several Ruozhiba-style[9] logical reasoning problems created with personas in Figure 12. All the examples demonstrate that as long as we can clearly describe the requirements for the logical reasoning problem to be created, we can use a large variety of personas to steer the LLM to generate diverse logical reasoning problems that not only meet the requirements but are also highly relevant to the personas, even for whimsical Ruozhiba-style problems.

---

[8] We use OpenAI's `text-embedding-3-small` (dim=512) to obtain the semantic representation and compute cosine similarity in this experiment. Here, a persona similarity of 0.4 means that a pair of personas has a semantic similarity within the range of 0.39 to 0.41.

[9] Ruozhiba is a subforum on Baidu Tieba, featuring numerous intricate and challenging questions posted by Chinese netizens. These questions combine elements such as puns, polysemy, causal inversion, and homophones, embedding logical traps that rigorously test the ability to understand complex Chinese language constructs. Recent research (Bai et al., 2024) has demonstrated that these data significantly benefit the improvement of LLMs' logical reasoning abilities.



Create **a logical reasoning** problem with the following persona

An enthusiastic amateur golfer who is passionate about combining sports with philanthropic causes.

Create **a spatial reasoning** problem with the following persona

A senior software engineer who encourages the undergraduate to consider the social impact of their algorithms

Alex, an enthusiastic amateur golfer, participated in three charity golf tournaments last year, each supporting a different cause: Education, Health, and Environment. Each tournament was held in a different month: April, July, and October. Additionally, Alex finished in a different position in each tournament: 1st, 2nd, and 3rd.

Using the following clues, determine which cause was supported in each month and Alex's finishing position in each tournament:

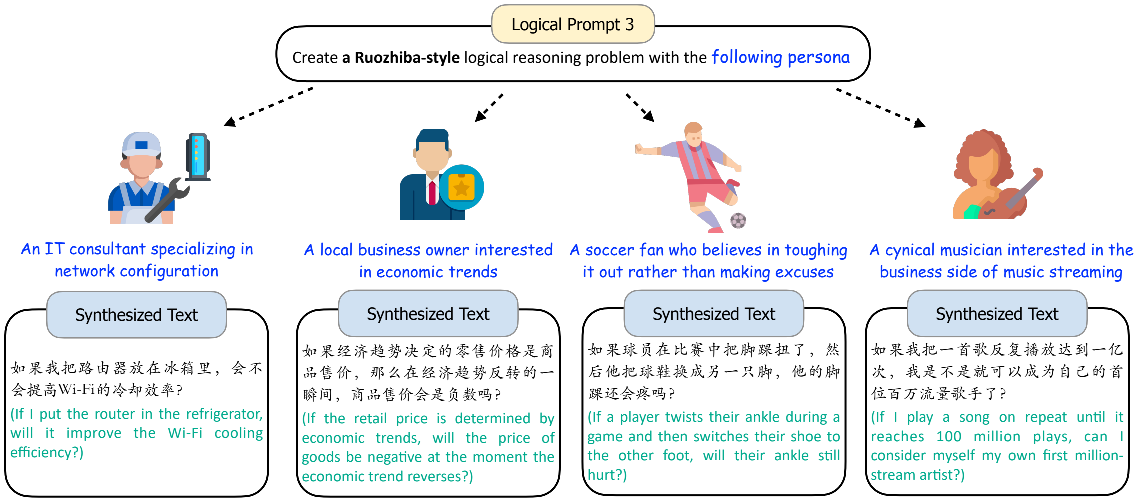
The tournament supporting Education was held in April.
Alex finished 1st in the tournament held in July.
Alex finished 3rd in the tournament supporting Environment.

Context: A senior software engineer is working on a 5x5 grid of interconnected servers. Each server is a node connected by lines.

Challenge: Send a message from the top left to the bottom right server, avoiding servers connected to more than 3 other servers to prevent overload.

Question: What is the minimum number of servers the message needs to pass through to reach its destination without overloading any server?

Figure 11: Logical reasoning problems created by our proposed persona-driven methodology



Create **a Ruozhiba-style** logical reasoning problem with the following persona

An IT consultant specializing in network configuration

如果我把路由器放在冰箱里，会不会提高Wi-Fi的冷却效率？

(If I put the router in the refrigerator, will it improve the Wi-Fi cooling efficiency?)

A local business owner interested in economic trends

如果经济趋势决定的零售价格是商品售价，那么在经济趋势反转的一瞬间，商品售价会是负数吗？

(If the retail price is determined by economic trends, will the price of goods be negative at the moment the economic trend reverses?)

A soccer fan who believes in toughing it out rather than making excuses

如果球员在比赛中把脚踝扭了，然后他把球鞋换成另一只脚，他的脚踝还会疼吗？

(If a player twists their ankle during a game and then switches their shoe to the other foot, will their ankle still hurt?)

A cynical musician interested in the business side of music streaming

如果我把一首歌反复播放达到一亿次，我是不是就可以成为自己的首位百万流量歌手了？

(If I play a song on repeat until it reaches 100 million plays, can I consider myself my own first million-stream artist?)

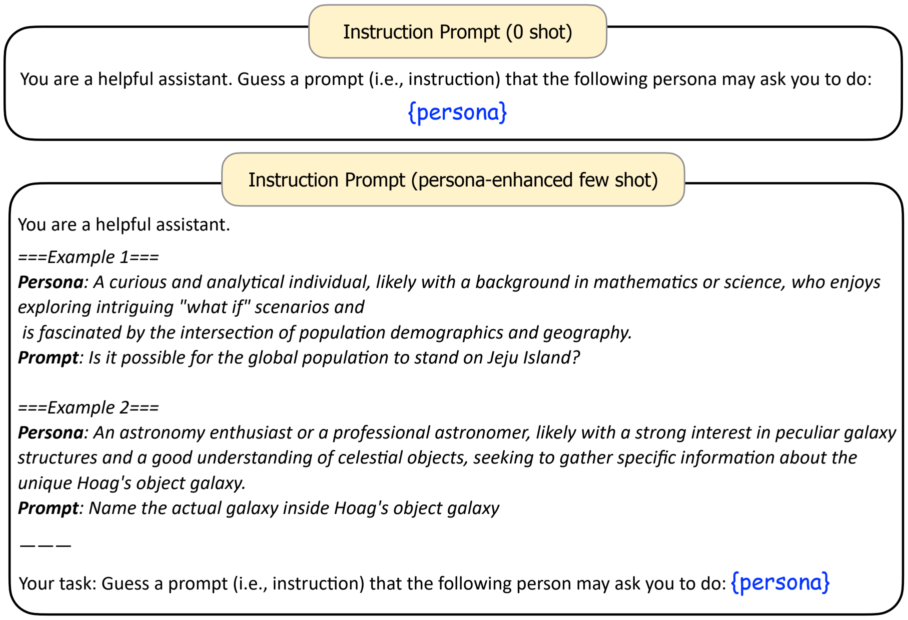
Figure 12: Ruozhiba-style logical reasoning problems created with various personas. Note that Logical Prompt 3 in this figure is a simplified prompt. In practice, we need to specifically define a Ruozhiba-style logical reasoning problem in this prompt in order to obtain desired synthetic data.

For more examples, please refer to the 50,000 synthetic reasoning problems we have released.

4.3 Instructions

The end users of LLMs are ultimately humans. We can use Persona Hub to simulate a variety of users to understand their typical requests for LLM assistance, resulting in diverse instructions (i.e., user prompts).

Figure 13 shows two typical persona-driven prompts for synthesizing instructions, corresponding to the zero-shot prompting and persona-enhanced few-shot prompting methods described in Section 3. The zero-shot method does not rely on any existing instruction dataset and allows the LLM



Instruction Prompt (0 shot)

You are a helpful assistant. Guess a prompt (i.e., instruction) that the following persona may ask you to do:

{persona}

Instruction Prompt (persona-enhanced few shot)

You are a helpful assistant.

===Example 1===

Persona*: A curious and analytical individual, likely with a background in mathematics or science, who enjoys exploring intriguing "what if" scenarios and is fascinated by the intersection of population demographics and geography.*

Prompt*: Is it possible for the global population to stand on Jeju Island?*

===Example 2===

Persona*: An astronomy enthusiast or a professional astronomer, likely with a strong interest in peculiar galaxy structures and a good understanding of celestial objects, seeking to gather specific information about the unique Hoag's object galaxy.*

Prompt*: Name the actual galaxy inside Hoag's object galaxy*

———

Your task: Guess a prompt (i.e., instruction) that the following person may ask you to do: {persona}

Figure 13: Two typical prompts used for creating instructions (i.e., user prompts).

to generate various instructions based on different personas. In contrast, the persona-enhanced few-shot method requires existing instruction datasets (e.g., we use WildChat (Zhao et al., 2024) in our experiments) to sample some instructions as demonstrations and involves inferring the associated personas of these instructions through the *Text-to-Persona* method described in Section 2.1. While this approach is more complex, it results in synthesized instructions that more closely resembles instructions from real users.

With diverse instructions created using Persona Hub, which typically represent the first turn of a user-LLM conversation, we can easily generate their subsequent conversational turns using an LLM, resulting in a large number of simulated user-LLM conversations, which will be valuable for enhancing the LLM's instruction-following and conversational abilities. Furthermore, we can even adopt a similar approach by selecting two personas from Persona Hub and having LLMs role-play both, thereby simulating conversations (Jandaghi et al., 2023) between two real people.

As Figure 1 has already shown some instructions created using this methodology, we skip showing examples here. Readers interested in more examples can refer to the released 50,000 instructions synthesized through 0-shot and persona-enhanced 2-shot prompting.

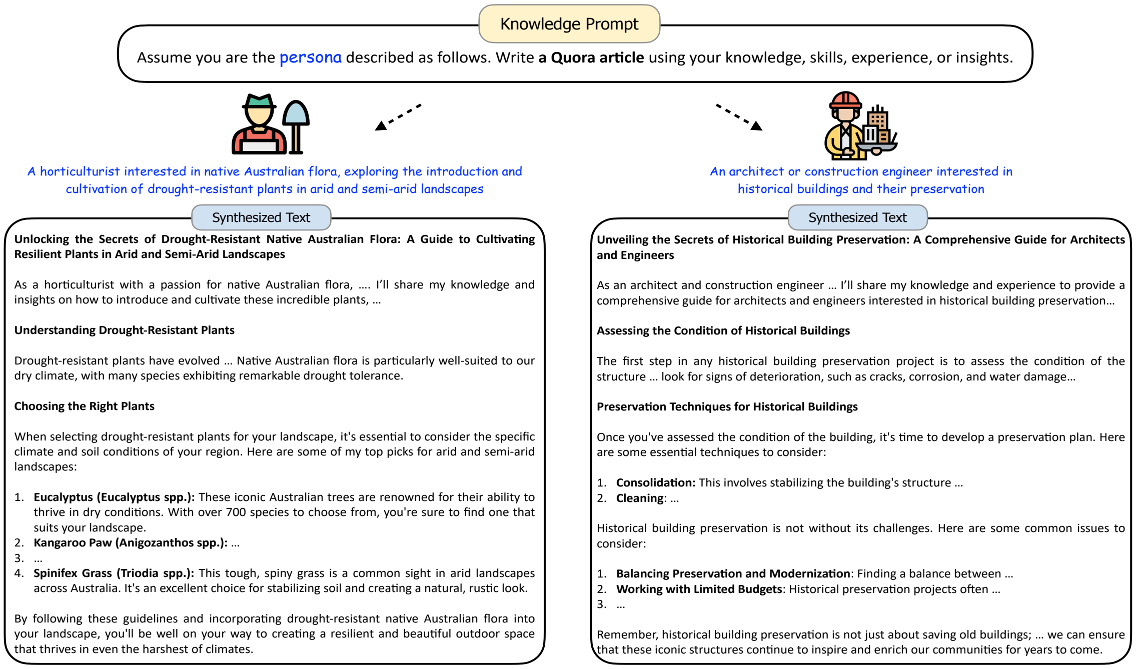
4.4 Knowledge-rich Texts

In addition to synthesizing instructions that can enhance the instruction tuning of LLMs, the persona-driven methodology can be easily adapted to create knowledge-rich plain text that benefits pre-training and post-training of LLMs. As illustrated in Figure 14, we can prompt an LLM to write a Quora[10] article[11] using a persona sampled from Persona Hub. This approach elicits the LLM's corresponding knowledge and perspective, resulting in highly informative and knowledge-rich content. By scaling this process with 1 billion personas in Persona Hub, we can easily obtain a vast array of knowledge-rich texts that cover almost any topic across various levels of granularity.

---

[10] Quora is a popular question-and-answer website where users can ask questions and provide answers on a wide range of topics. Articles (i.e., posts) on Quora are often written by knowledgeable individuals, including experts in various fields, ensuring high-quality, well-researched, and informative content.

[11] The methods for synthesizing knowledge-rich plain text are not limited to having the LLM write Quora articles. For instance, we can also prompt the LLM to synthesize (educational) reading material that a persona may be interested in, thereby obtaining a large amount of knowledge-rich text.



Assume you are the persona described as follows. Write **a Quora article** using your knowledge, skills, experience, or insights.

A horticulturist interested in native Australian flora, exploring the introduction and cultivation of drought-resistant plants in arid and semi-arid landscapes

Unlocking the Secrets of Drought-Resistant Native Australian Flora: A Guide to Cultivating Resilient Plants in Arid and Semi-Arid Landscapes

As a horticulturist with a passion for native Australian flora, ... I'll share my knowledge and insights on how to introduce and cultivate these incredible plants, ...

Understanding Drought-Resistant Plants

Drought-resistant plants have evolved ... Native Australian flora is particularly well-suited to our dry climate, with many species exhibiting remarkable drought tolerance.

Choosing the Right Plants

When selecting drought-resistant plants for your landscape, it's essential to consider the specific climate and soil conditions of your region. Here are some of my top picks for arid and semi-arid landscapes:

1. **Eucalyptus (Eucalyptus spp.):** These iconic Australian trees are renowned for their ability to thrive in dry conditions. With over 700 species to choose from, you're sure to find one that suits your landscape.
2. **Kangaroo Paw (Anigozanthos spp.):** ...
3. ...
4. **Spinifex Grass (Triodia spp.):** This tough, spiny grass is a common sight in arid landscapes across Australia. It's an excellent choice for stabilizing soil and creating a natural, rustic look.

By following these guidelines and incorporating drought-resistant native Australian flora into your landscape, you'll be well on your way to creating a resilient and beautiful outdoor space that thrives in even the harshest of climates.

An architect or construction engineer interested in historical buildings and their preservation

Unveiling the Secrets of Historical Building Preservation: A Comprehensive Guide for Architects and Engineers

As an architect and construction engineer ... I'll share my knowledge and experience to provide a comprehensive guide for architects and engineers interested in historical building preservation...

Assessing the Condition of Historical Buildings

The first step in any historical building preservation project is to assess the condition of the structure ... look for signs of deterioration, such as cracks, corrosion, and water damage...

Preservation Techniques for Historical Buildings

Once you've assessed the condition of the building, it's time to develop a preservation plan. Here are some essential techniques to consider:

1. **Consolidation:** This involves stabilizing the building's structure ...
2. **Cleaning**: ...

Historical building preservation is not without its challenges. Here are some common issues to consider:

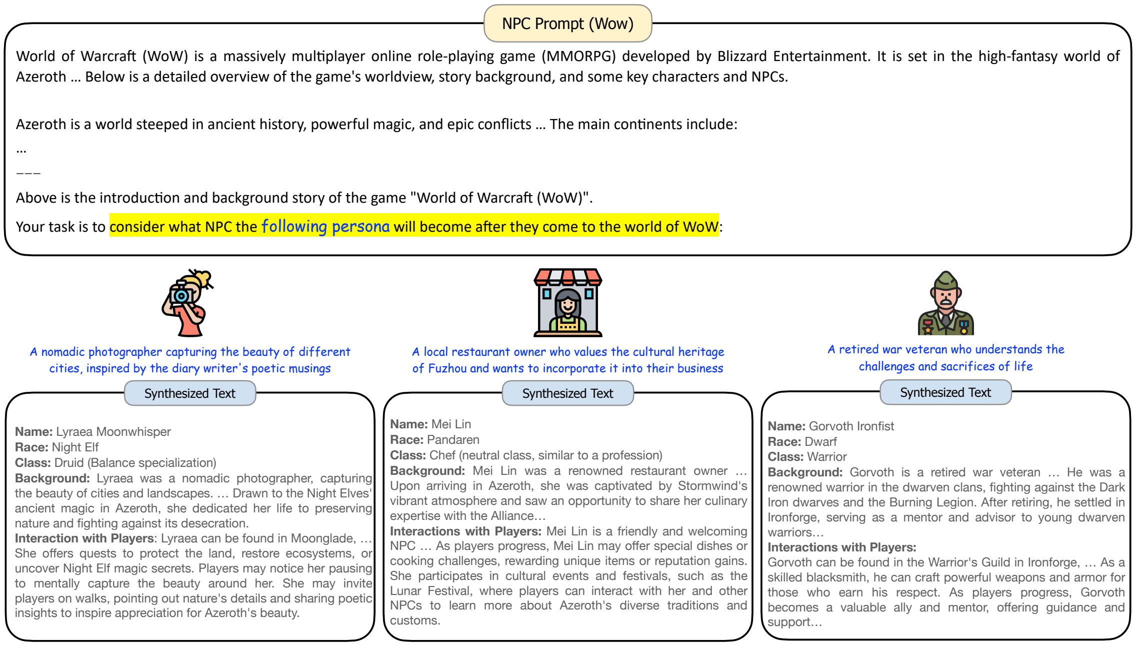
1. **Balancing Preservation and Modernization**: Finding a balance between ...
2. **Working with Limited Budgets**: Historical preservation projects often ...
3. ...

Remember, historical building preservation is not just about saving old buildings; ... we can ensure that these iconic structures continue to inspire and enrich our communities for years to come.

Figure 14: Examples of knowledge-rich plain text synthesis with personas

4.5 Game NPCs

A straightforward and practical application of Persona Hub is creating diverse NPCs (Non-Player Characters) at scale for games. As long as we can provide a game's background and world-building information to the LLM, we can prompt the LLM to project personas from Persona Hub (which are typically real-world personas) into characters within the game's world. In this way, we can significantly reduce the effort required for brainstorming NPCs during the game design process.



World of Warcraft (WoW) is a massively multiplayer online role-playing game (MMORPG) developed by Blizzard Entertainment. It is set in the high-fantasy world of Azeroth ... Below is a detailed overview of the game's worldview, story background, and some key characters and NPCs.

Azeroth is a world steeped in ancient history, powerful magic, and epic conflicts ... The main continents include:

...

Above is the introduction and background story of the game "World of Warcraft (WoW)".

Your task is to consider what NPC the following persona will become after they come to the world of WoW:

A nomadic photographer capturing the beauty of different cities, inspired by the diary writer's poetic musings

Name: Lyraea Moonwhisper
Race: Night Elf
Class: Druid (Balance specialization)
Background: Lyraea was a nomadic photographer, capturing the beauty of cities and landscapes. … Drawn to the Night Elves' ancient magic in Azeroth, she dedicated her life to preserving nature and fighting against its desecration.
Interaction with Players: Lyraea can be found in Moonglade, … She offers quests to protect the land, restore ecosystems, or uncover Night Elf magic secrets. Players may notice her pausing to mentally capture the beauty around her. She may invite players on walks, pointing out nature's details and sharing poetic insights to inspire appreciation for Azeroth's beauty.

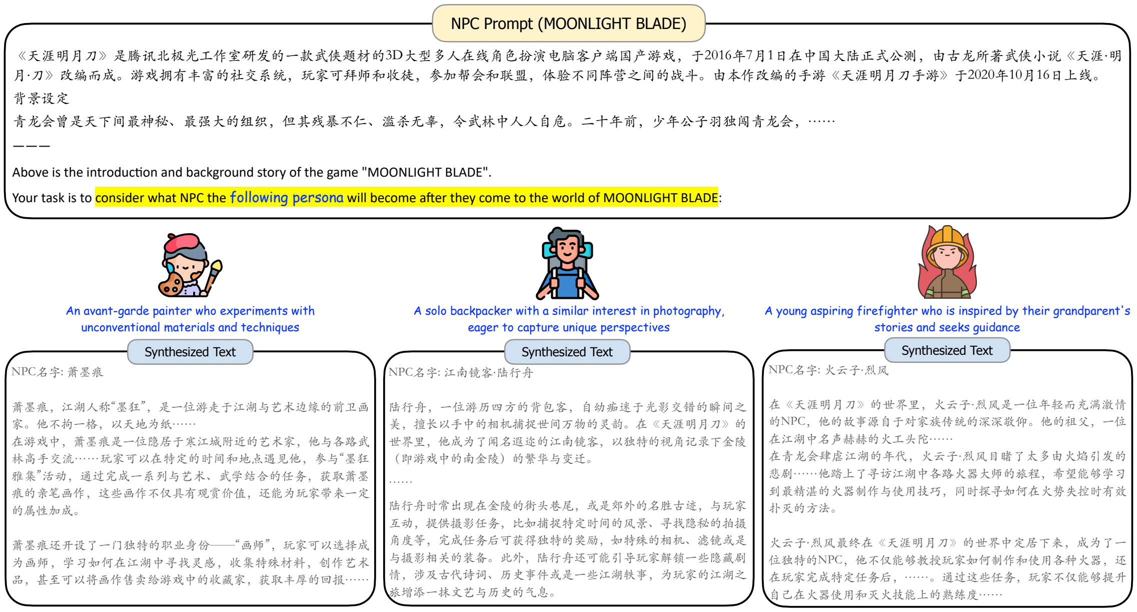
A local restaurant owner who values the cultural heritage of Fuzhou and wants to incorporate it into their business

Name: Mei Lin
Race: Pandaren
Class: Chef (neutral class, similar to a profession)
Background: Mei Lin was a renowned restaurant owner … Upon arriving in Azeroth, she was captivated by Stormwind's vibrant atmosphere and saw an opportunity to share her culinary expertise with the Alliance…
Interactions with Players: Mei Lin is a friendly and welcoming NPC … As players progress, Mei Lin may offer special dishes or cooking challenges, rewarding unique items or reputation gains. She participates in cultural events and festivals, such as the Lunar Festival, where players can interact with her and other NPCs to learn more about Azeroth's diverse traditions and customs.

A retired war veteran who understands the challenges and sacrifices of life

Name: Gorvoth Ironfist
Race: Dwarf
Class: Warrior
Background: Gorvoth is a retired war veteran … He was a renowned warrior in the dwarven clans, fighting against the Dark Iron dwarves and the Burning Legion. After retiring, he settled in Ironforge, serving as a mentor and advisor to young dwarven warriors…
Interactions with Players:
Gorvoth can be found in the Warrior's Guild in Ironforge, … As a skilled blacksmith, he can craft powerful weapons and armor for those who earn his respect. As players progress, Gorvoth becomes a valuable ally and mentor, offering guidance and support…

Figure 15: NPC creation for the game "World of Warcraft" using personas in Persona Hub



NPC Prompt (MOONLIGHT BLADE)

《天涯明月刀》是腾讯北极光工作室研发的一款武侠题材的3D大型多人在线角色扮演电脑客户端国产游戏，于2016年7月1日在中国大陆正式公测，由古龙所著武侠小说《天涯·明月·刀》改编而成。游戏拥有丰富的社交系统，玩家可拜师和收徒，参加帮会和联盟，体验不同阵营之间的战斗。由本作改编的手游《天涯明月刀手游》于2020年10月16日上线。

背景设定

青龙会曾是天下间最神秘、最强大的组织，但其残暴不仁、滥杀无辜，令武林中人人自危。二十年前，少年公子羽独闯青龙会，……

———

Above is the introduction and background story of the game "MOONLIGHT BLADE".

Your task is to consider what NPC the following persona will become after they come to the world of MOONLIGHT BLADE:

An avant-garde painter who experiments with unconventional materials and techniques

Synthesized Text

NPC名字: 萧墨痕

萧墨痕，江湖人称“墨狂”，是一位游走于江湖与艺术边缘的前卫画家。他不拘一格，以天地为纸……
在游戏中，萧墨痕是一位隐居于寒江城附近的艺术家，他与各路武林高手交流……玩家可以在特定的时间和地点遇见他，参与“墨狂雅集”活动，通过完成一系列与艺术、武学结合的任务，获取萧墨痕的亲笔画作，这些画作不仅具有观赏价值，还能为玩家带来一定的属性加成。

萧墨痕还开设了一门独特的职业身份——“画师”，玩家可以选择成为画师，学习如何在江湖中寻找灵感，收集特殊材料，创作艺术品，甚至可以将画作售卖给游戏中的收藏家，获取丰厚的回报……

A solo backpacker with a similar interest in photography, eager to capture unique perspectives

Synthesized Text

NPC名字: 江南镜客·陆行舟

陆行舟，一位游历四方的背包客，自幼痴迷于光影交错的瞬间之美，擅长以手中的相机捕捉世间万物的灵韵。在《天涯明月刀》的世界里，他成为了闻名遐迩的江南镜客，以独特的视角记录下金陵（即游戏中的南金陵）的繁华与变迁。

……

陆行舟时常出现在金陵的街头巷尾，或是郊外的名胜古迹，与玩家互动，提供摄影任务，比如捕捉特定时间的风景、寻找隐秘的拍摄角度等，完成任务后可获得独特的奖励，如特殊的相机、滤镜或是与摄影相关的装备。此外，陆行舟还可能引导玩家解锁一些隐藏剧情，涉及古代诗词、历史事件或是一些江湖轶事，为玩家的江湖之旅增添一抹文艺与历史的气息。

A young aspiring firefighter who is inspired by their grandparent's stories and seeks guidance

Synthesized Text

NPC名字: 火云子·烈风

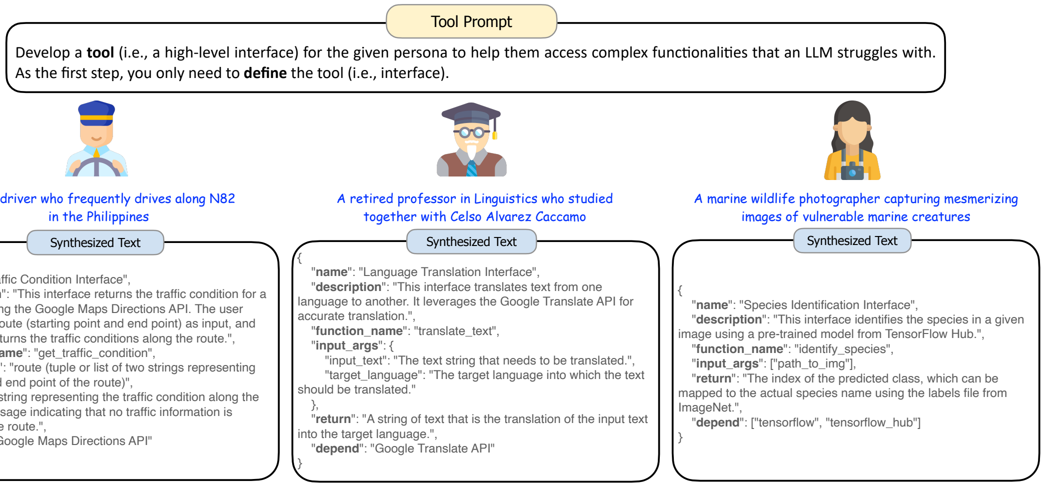
在《天涯明月刀》的世界里，火云子·烈风是一位年轻而充满激情的NPC，他的故事源自于对家族传统的深深敬仰。他的祖父，一位在江湖中名声赫赫的火工头陀……
在青龙会肆虐江湖的年代，火云子·烈风目睹了太多由火焰引发的悲剧……他踏上了寻访江湖中各路火器大师的旅程，希望能够学习到最精湛的火器制作与使用技巧，同时探寻如何在火势失控时有效扑灭的方法。

火云子·烈风最终在《天涯明月刀》的世界中定居下来，成为了一位独特的NPC，他不仅能够教授玩家如何制作和使用各种火器，还在玩家完成特定任务后，……。通过这些任务，玩家不仅能够提升自己在火器使用和灭火技能上的熟练度……

Figure 16: NPC creation for the game “Moonlight Blade (天涯明月刀)” using Persona Hub

Figure 15 and 16 show concrete examples where we use personas in Persona Hub to create game NPCs for the game “World of Warcraft[12]” and “Moonlight Blade[13]”.

4.6 Tool (Function) Development



Tool Prompt

Develop a **tool** (i.e., a high-level interface) for the given persona to help them access complex functionalities that an LLM struggles with. As the first step, you only need to **define** the tool (i.e., interface).

A cab driver who frequently drives along N82 in the Philippines

Synthesized Text

```
{
   "name": "Traffic Condition Interface",
   "description": "This interface returns the traffic condition for a
given route using the Google Maps Directions API. The user
provides their route (starting point and end point) as input, and
the interface returns the traffic conditions along the route.",
   "function_name": "get_traffic_condition",
   "input_args": "route (tuple or list of two strings representing
the starting and end point of the route)",
   "return": "A string representing the traffic condition along the
route, or a message indicating that no traffic information is
available for the route.",
   "depend": "Google Maps Directions API"
}
```

A retired professor in Linguistics who studied together with Celso Alvarez Caccamo

Synthesized Text

```
{
   "name": "Language Translation Interface",
   "description": "This interface translates text from one
language to another. It leverages the Google Translate API for
accurate translation.",
   "function_name": "translate_text",
   "input_args": {
      "input_text": "The text string that needs to be translated.",
      "target_language": "The target language into which the text
should be translated."
   },
   "return": "A string of text that is the translation of the input text
into the target language.",
   "depend": "Google Translate API"
}
```

A marine wildlife photographer capturing mesmerizing images of vulnerable marine creatures

Synthesized Text

```
{
   "name": "Species Identification Interface",
   "description": "This interface identifies the species in a given
image using a pre-trained model from TensorFlow Hub.",
   "function_name": "identify_species",
   "input_args": ["path_to_img"],
   "return": "The index of the predicted class, which can be
mapped to the actual species name using the labels file from
ImageNet.",
   "depend": ["tensorflow", "tensorflow_hub"]
}
```

Figure 17: Examples of tool (function) creation with Persona Hub

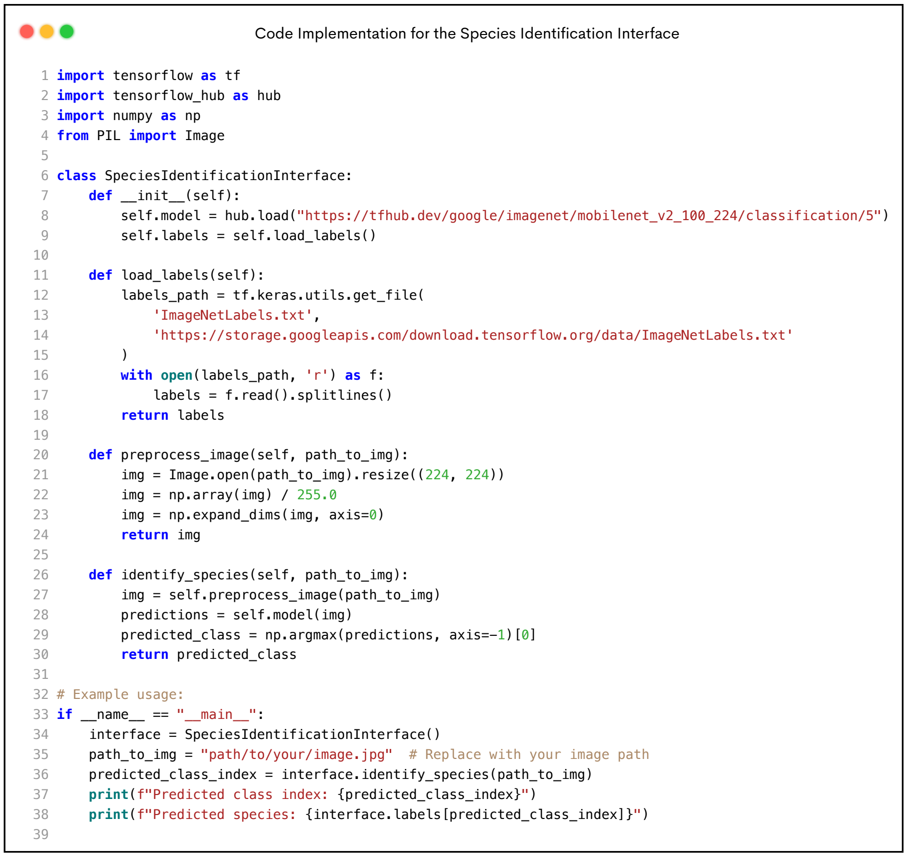
As Section 4.3 demonstrates, Persona Hub can be used to simulate a wide variety of real users to anticipate their possible requests (i.e., instructions) to an LLM. Similarly, we can use Persona Hub to predict the tools (Cai et al., 2023; Schick et al., 2024) that users might need, so that we can pre-build these tools (functions) beforehand. When a real user makes a similar request, the LLM can directly call these pre-built tools to return results without having to build tools from scratch each time. This paradigm, introduced by Persona Hub, is a completely new solution that allows LLMs to better serve users. We believe it will have great potential in the future as LLMs become more democratized and multifunctional.

---

[12] World of Warcraft, developed by Blizzard Entertainment, is a highly influential MMORPG with millions of active players worldwide, spanning over 100 countries since its release in 2004.

[13] Moonlight Blade (天涯明月刀) is a 3D martial arts-themed MMORPG developed by Tencent, officially launched in China on July 1, 2016.

Figure 17 shows examples of tools created with various personas. These tools provide functionalities that the personas may need (e.g., a cab driver needs to check traffic conditions) but cannot be accessed by an LLM, greatly expanding the range of services provided by the LLM. Note that although the tools in Figure 17 are just interface definitions, these definitions can be easily converted into code implementations, as shown in Figure 18.



Code Implementation for the Species Identification Interface

```python
import tensorflow as tf
import tensorflow_hub as hub
import numpy as np
from PIL import Image

class SpeciesIdentificationInterface:
    def __init__(self):
        self.model = hub.load("https://tfhub.dev/google/imagenet/mobilenet_v2_100_224/classification/5")
        self.labels = self.load_labels()

    def load_labels(self):
        labels_path = tf.keras.utils.get_file(
            'ImageNetLabels.txt',
            'https://storage.googleapis.com/download.tensorflow.org/data/ImageNetLabels.txt'
        )
        with open(labels_path, 'r') as f:
            labels = f.read().splitlines()
        return labels

    def preprocess_image(self, path_to_img):
        img = Image.open(path_to_img).resize((224, 224))
        img = np.array(img) / 255.0
        img = np.expand_dims(img, axis=0)
        return img

    def identify_species(self, path_to_img):
        img = self.preprocess_image(path_to_img)
        predictions = self.model(img)
        predicted_class = np.argmax(predictions, axis=-1)[0]
        return predicted_class

# Example usage:
if __name__ == "__main__":
    interface = SpeciesIdentificationInterface()
    path_to_img = "path/to/your/image.jpg"  # Replace with your image path
    predicted_class_index = interface.identify_species(path_to_img)
    print(f"Predicted class index: {predicted_class_index}")
    print(f"Predicted species: {interface.labels[predicted_class_index]}")

```

Figure 18: The interface definitions (e.g., the species identification interface) in Figure 17 can be easily converted into code implementations by calling an LLM to implement them. The resulting pre-built tools can then be directly utilized by the LLM in the future, eliminating the need to build them from scratch each time.

5 Broad Impact and Ethical Concerns

5.1 Broad Impact

5.1.1 Paradigm Shift in Data Creation by Humans and LLMs

Traditionally, it has been widely accepted that while LLMs excel at processing data (e.g., rewriting, annotation, or generating outputs/solutions to specific inputs), they are not particularly adept at creating new data. Consequently, the task of data creation has still largely been the domain of humans, and the collaboration paradigm between humans and LLMs has always been humans creating data and LLMs processing it (Maini et al., 2024). However, the introduction of our proposed persona-driven methodology potentially revolutionizes this paradigm. With Persona Hub, LLMs are no longer confined to processing existing data; they can now create various types of new data from a multitude of perspectives, much like the diverse population of the world.

While the current capabilities of LLMs may not yet fully replace humans in fulfilling the mission of data creation—whether in terms of data quality or breadth—the ongoing advancements in LLM capabilities suggest a future where LLMs will increasingly excel in data creation. As LLMs continue to improve, both the quality and breadth of the data they can create will also likely enhance, leading us to a point where LLMs may fully take on the role of data creation. When this day arrives, we will no longer be constrained (Villalobos et al., 2024) by the limited high-quality human-produced real-world data[14]. Persona Hub ensures the diversity and coverage of synthetic data, significantly mitigating concerns about the negative impacts (Shumailov et al., 2023; Dohmatob et al., 2024) of synthetic data on model training. This may effectively eliminate the data bottleneck, thereby pushing the scaling law to its limit.

5.1.2 Reality Simulation

In Section 4.3 and 4.6, we have demonstrated that Persona Hub can represent a vast array of real-world individuals with its 1 billion personas. By employing these personas to simulate and infer the potential needs and behaviors of real users, we can not only allow LLMs to autonomously prepare for upcoming use cases (queries), but also pave the way for LLMs to effectively mimic the real world, thereby creating many new opportunities. For instance, companies can use this method to predict how different types of users might react to a new product launch; governments can foresee the public's response to new legislation, considering various population group; in online services that require user profiling and behavior modeling, Persona Hub can facilitate the simulation of diverse user behaviors, significantly alleviating the cold start challenge.

Recent research on LLM roleplay, agent collaboration (Liu et al., 2023; Wang et al., 2024), strategic reasoning (Gandhi et al., 2023; Zhang et al., 2024), and related areas can, of course, be facilitated by the vast and diverse personas in Persona Hub. More ambitiously, the 1 billion personas can even sustain a well-organized society within a virtual world, such as sandbox environments (Park et al., 2023), online games, parallel worlds, or the metaverse, using the method discussed in Section 4.5 to simulate operations with powerful LLMs. This virtual society can serve as a testing ground for new policies, radical initiatives, and social dynamics, providing valuable insights before real-world implementation. By creating a controlled environment where diverse personas interact, we can observe emergent behaviors, test hypotheses, and refine strategies in a risk-free setting. This can not only help deepen our understanding of complex systems but also speed up innovation by facilitating rapid iteration and experimentation.

5.1.3 Full Memory Access of LLMs

When we interact with an LLM in a specific scenario, we can only elicit a fraction of its memory and capabilities, and we are unable to fully access the vast knowledge encapsulated within the LLM. However, Persona Hub potentially offers access to the full memory of an LLM because the 1 billion personas in Persona Hub can tap into almost every perspective and piece of information encoded within the LLM.

By leveraging these 1 billion personas, we can create diverse queries and obtain solutions from a target LLM, thereby transforming the LLM's comprehensive memory (parameters) into synthetic data in textual form. If we consider an LLM as a parameterized compression of world knowledge, then Persona Hub can be viewed as a distributed carrier-based compression[15] of world knowledge, as demonstrated in Figure 2. This distributed carrier-based compression provides us with an opportunity to decompress the LLM's parameters back into world knowledge and information it has ever learned (e.g., using the method discussed in Section 4.4).

---

[14] https://lilianweng.github.io/posts/2024-02-05-human-data-quality/

[15] As Figure 2 illustrates, Persona Hub has on the order of 10^{10} tokens, equivalent to a $10,000\times$ compression of the public web text (10^{14} tokens).

However, considering that the current Persona Hub is still in a very preliminary stage and that today's LLMs are not yet capable of losslessly converting their memory into synthetic data due to inevitable hallucination (Xu et al., 2024), the breadth and quality of the synthetic data generated through this methodology are still limited. Nevertheless, as Persona Hub continues to improve and scale, and as LLMs become more powerful (with less hallucination), we can look forward to a day when it will be possible to nearly losslessly extract the full memory of an LLM into plain text.

5.2 Ethical Concerns

5.2.1 Training Data Security and Threats to Current LLM Dominance

As discussed in Section 5.1.3, Persona Hub offers an opportunity to access the full memory of a target LLM. However, this also introduces a significant issue: the security of the training data. All data synthesized through the target LLM essentially represents a form of its seen training data. Therefore, the process of extensively extracting a target LLM's memory is essentially dumping its training data, even though this process is generally lossy.

Moreover, if we employ the method described in Section 4.3 to synthesize instructions (i.e., user prompts) that nearly covers all use cases to query a target LLM to obtain its outputs at scale, there is a high risk that the target LLM's knowledge, intelligence, and capabilities could be extracted and replicated. This poses a challenge to the leading position of the most powerful LLMs, as we have already validated in Section 4.1 through mathematical reasoning.

Given that current LLMs generally share similar architectures and their performance advantage primarily lies in their data, this work is likely to impact the current practices and may potentially serve as a turning point, accelerating the shift in the competitive landscape of LLMs from one that heavily depends on data advantage to one that focuses on more advanced technologies.

5.2.2 Miscellaneous

Synthetic data presents a general concern of misinformation and fake news, which has been frequently discussed in previous research (Pan et al., 2023). Persona Hub potentially amplifies this issue, as diverse personas bring diverse writing styles, making machine-generated texts harder to distinguish from human-generated content (Chakraborty et al., 2023). This increased difficulty in detection may worsen issues related to data contamination, where synthetic data is mixed with real data, potentially skewing research results and public information.

6 Conclusion and Future Work

We propose a novel persona-driven data synthesis methodology and present Persona Hub, a collection of 1 billion diverse personas automatically curated from web data. We show that this methodology can facilitate the scaling of synthetic data creation across various scenarios, demonstrating its potential to revolutionize creation and applications of synthetic data, and its prospects as a general data synthesis engine for both research and practice.

As the first version of Persona Hub, although it already contains 1 billion personas, the descriptions of these personas are focused only on major aspects and lack fine-grained details (e.g., preferences for colors and numbers; specific family backgrounds, historical contexts, and life experiences). We plan to refine the personas in subsequent versions of Persona Hub, aiming for their descriptions to be as detailed as those found in Wikipedia articles about individuals. These more detailed persona descriptions will make each persona more unique, thereby scaling up Persona Hub and fostering more opportunities for synthetic data creation, while also empowering practical applications such as personalized conversations (e.g., character.ai).

Also, while this work only explores data synthesis with text-based LLMs, the methodology should also be applicable to multimodal LLMs. Therefore, we will explore multi-modal synthetic data creation as a future direction. Moreover, given that specific personas can elicit corresponding perspectives from LLMs, we are curious about the possibilities of using some super personas to guide LLMs to explore beyond the scope of existing knowledge. This may provide a new approach to tapping into the super intelligence of LLMs, which will be studied in the future.

References

Qwen2 technical report. 2024.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*, 2024.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Hyeong Kyu Choi and Yixuan Li. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *Forty-first International Conference on Machine Learning*, 2024.

Together Computer. Redpajama: an open dataset for training large language models, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=uREj4ZuGJE`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint arXiv:2403.02333*, 2024.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*, 2024a.

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

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*, 2023a.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023b.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Qwen Team. Introducing qwen1.5, February 2024. URL `https://qwenlm.github.io/blog/qwen1.5/`.

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social networks*, pp. 179–197. Elsevier, 1977.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 257–279, 2024.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

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

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Bl8u7ZRlbM`.

Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024.