Section 4.1: Generative Models for **Synthetic Data Generation (SDG)**

2026 Spring

LLM Agents Foundation & Applications

Dr. Yanjun Qi
20260220

Last Class

- LLM-based Agents in Health Care:
  - 2026-SP-W3.1-HCLS-agent.pdf
  - 2026-SP-W3.2-Team05-Agent-Healthcare.pdf
- LLM-based Agents in Life Science:
  - 2026-SP-W3.3-Team06-BioinfomaticsAgents.pdf
  - 2026-SP-W3.4-Team0506-HCLS-agentData
- Foundation Models in HCLS
  - 2026-SP-W3.6-HealthCare-FMs.pdf
  - 2026-SP-W3.5-LifeScience-ProteinLLM.pdf
  - 2026-SP-W3.7-LifeScience-genomeLLM

This class: Reference

- 1. Guo & Chen (2024). "Generative AI for Synthetic Data Generation: Methods, Challenges and the Future”
- 2. Constitutional AI: Harmlessness from AI Feedback (2022)
- 3. SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions (2022)
- 4. Scaling Synthetic Data Creation with 1M Personas (2024)
- 5. Bauer et al. (2024). "Comprehensive Exploration of Synthetic Data Generation: A Survey."
- 6. GigaTIME: Multimodal AI Generates a Virtual Population for Tumor Microenvironment Modeling (**Cell, January 2026**)

Basics of SDG

What is Synthetic Data Generation (SDG)?

- A generative model **learns the statistical distribution** of real data, then samples brand-new artificial examples
- Generated data is **novel** — not a copy or a transformation of any real record
- It can be produced at **any scale**, for **any class**, under **any constraint**

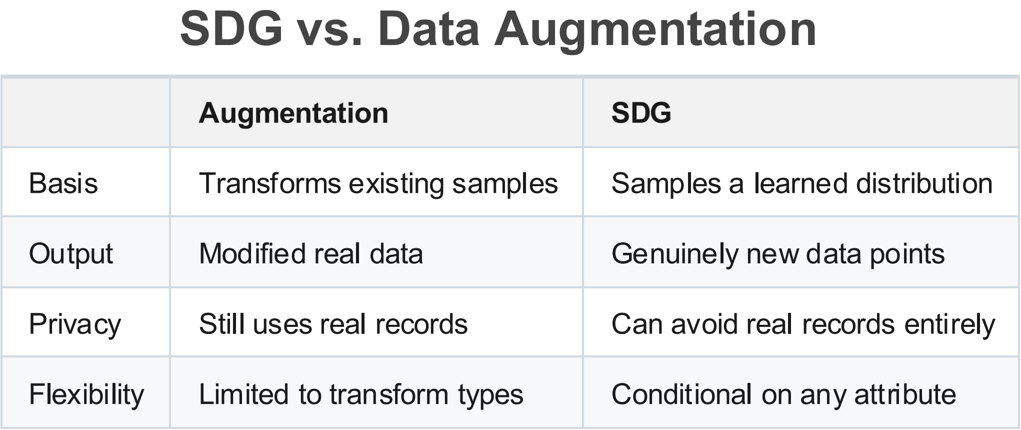
The Core Problem: Real Data is Not Enough

- Privacy & Regulation — GDPR and sensitive PII restrict sharing medical, financial, or biometric data
- Scarcity & Cost — labeling real data is expensive; minority classes are underrepresented
- Bias & Imbalance — real-world datasets reflect historical biases that harm model fairness

Solution: Synthetic Data Generation (SDG) produces unlimited, pre-labeled, privacy-safe samples

SDG is fundamentally different from data augmentation

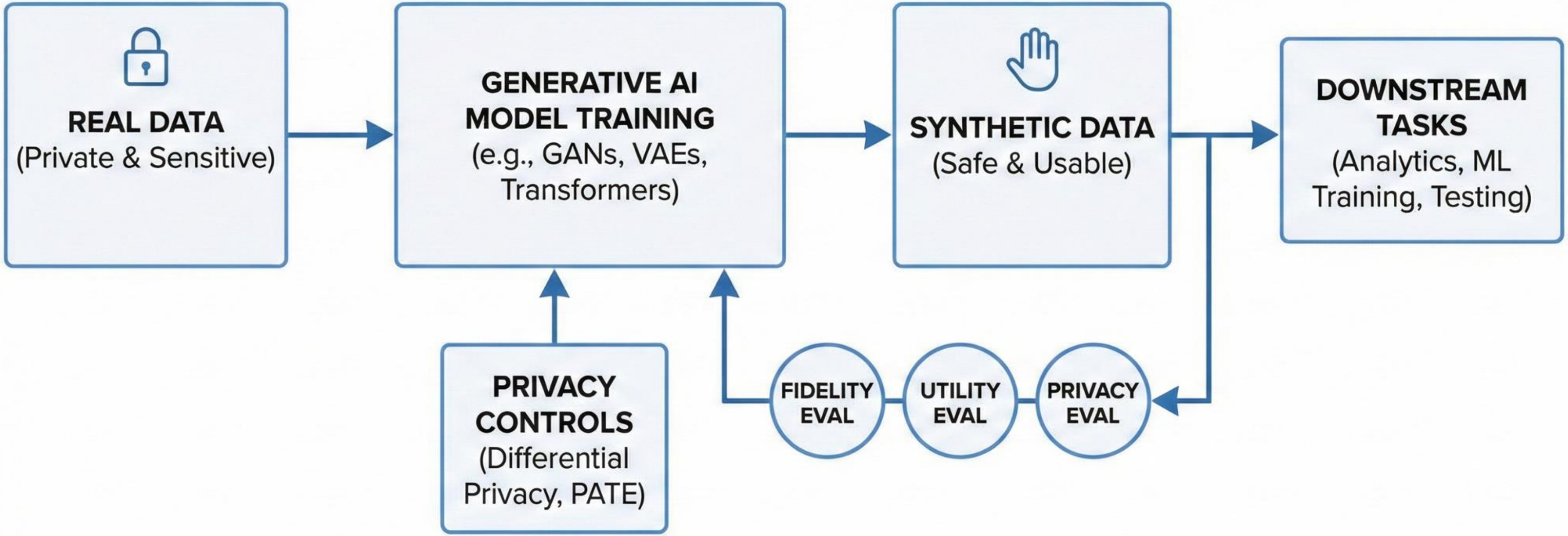
- Data augmentation transforms existing, samples.
- SDG samples from a learned distribution to create genuinely new data.



SDG vs. Data Augmentation

	Augmentation	SDG
Basis	Transforms existing samples	Samples a learned distribution
Output	Modified real data	Genuinely new data points
Privacy	Still uses real records	Can avoid real records entirely
Flexibility	Limited to transform types	Conditional on any attribute

SYNTHETIC DATA GENERATION PIPELINE



REAL DATA
(Private & Sensitive)

GENERATIVE AI MODEL TRAINING
(e.g., GANs, VAEs, Transformers)

SYNTHETIC DATA
(Safe & Usable)

DOWNSTREAM TASKS
(Analytics, ML Training, Testing)

PRIVACY CONTROLS
(Differential Privacy, PATE)

FIDELITY EVAL

UTILITY EVAL

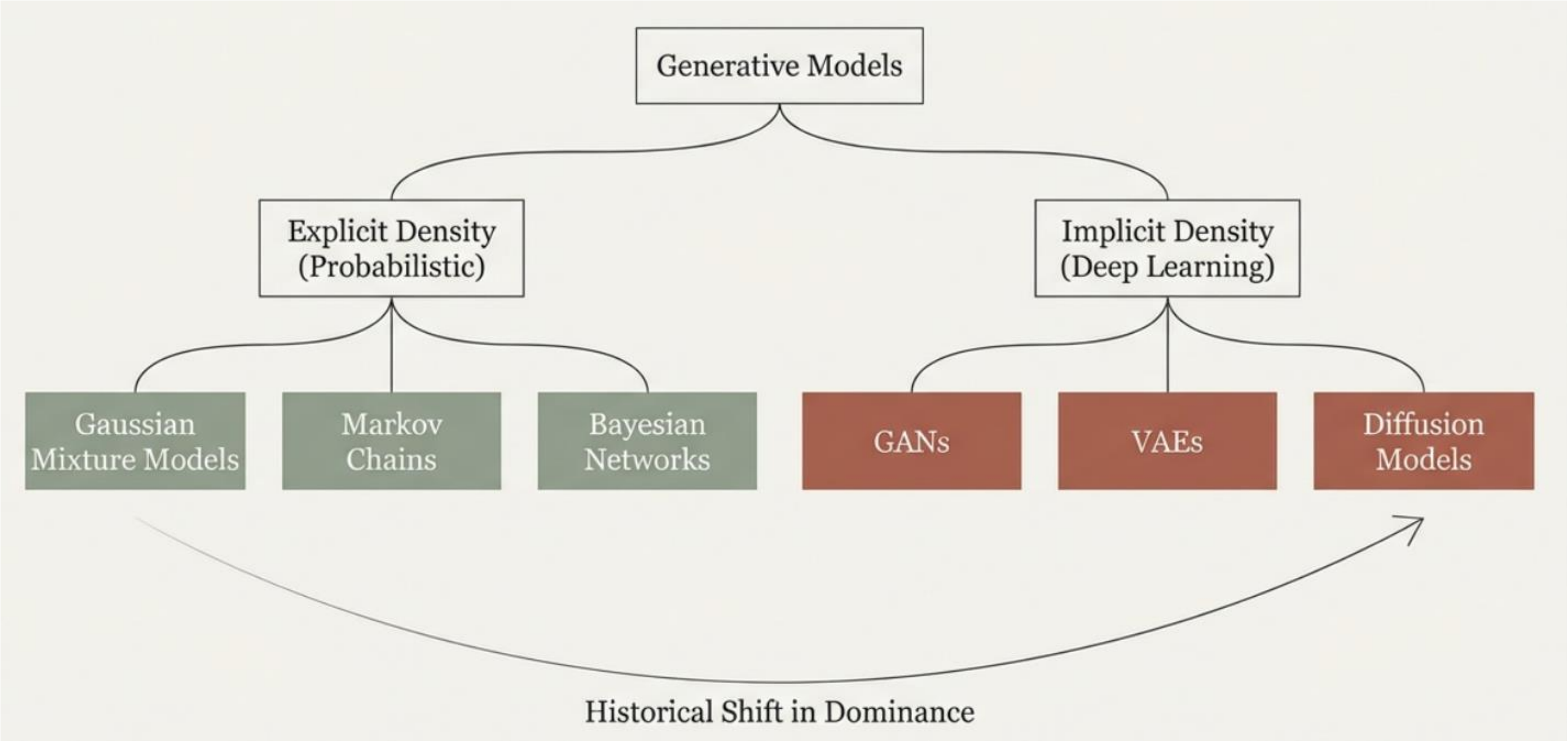
PRIVACY EVAL

Data Modalities & Challenges

- **Tabular / Relational**: Structured records are difficult due to mixed data types, missingness, high-cardinality, and multi-table rules.
- **Time-Series**: Sequences with temporal correlation face challenges with irregular sampling and preserving event ordering.
- **Text**: Documents and clinical notes have a high risk of memorizing rare text spans like names or addresses.
- **Image / Video**: Perception stacks require perceptual realism and strict label fidelity.
- **Multimodal**: Sensor fusion or image+text requires cross-modal alignment.

Generative Models Based SDG

Bauer et al. (2024). "Comprehensive Exploration of Synthetic Data Generation: A Survey.



417 Synthetic Data Generation (SDG) models over the last decade, providing a comprehensive overview of model types, functionality, and improvements.

Literature’s 20 distinct model families for SDG

- **Classical / probabilistic** — Gaussian Mixtures, Markov Chains, Bayesian Networks
- **Energy-based** — Boltzmann Machines
- **Neural generative** — VAEs, Normalizing Flows, GANs, Diffusion Models
- **Sequential / attention** — RNNs (LSTM/GRU), Transformers
- **Other** — RL-guided generation, simulation-based virtual environments

> The survey covers all 20 families (42 subtypes) across 417 papers from 2012–2022 — giving practitioners the most comprehensive map of the SDG space to date.

Why Classical Models Failed to Scale

- Gaussian Mixture Models fit well to simple tabular data but cannot capture high-dimensional image structure
- Markov Chains produce realistic sequences but suffer from 'amnesia' — no long-term memory
- Bayesian Networks encode causal relationships but scale poorly to thousands of variables
- The gap: none of these can synthesize a photorealistic image from a text description

Key Recent Generative Model Families

- **GANs**: Provide sharp samples and flexible conditioning, but suffer from training instability and mode collapse.
- **VAEs**: Offer stable training and explicit latent structure, though often produce blurrier images compared to GANs.
- **Diffusion Models**: Deliver strong sample quality and good mode coverage, but require intensive computing power.
- **Autoregressive Transformers**: Yield strong semantic coherence for text, but carry high risks of verbatim memorization and PII leakage.

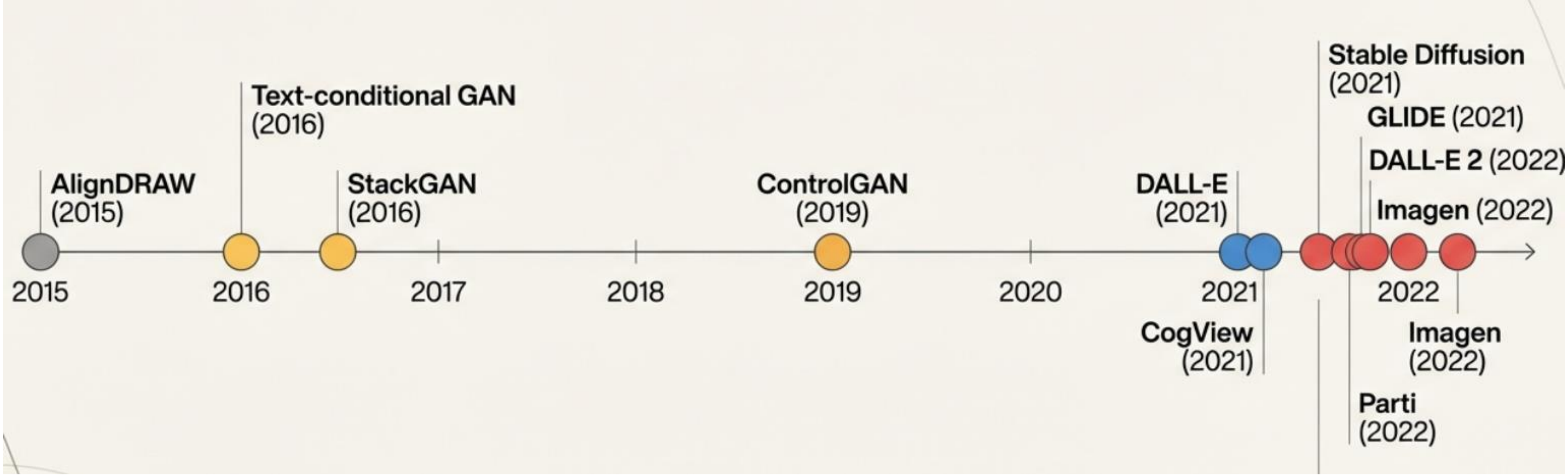
The right model depends on your data type

Data type	Best-fit model families
Images	GANs, Diffusion models, VAEs
Text / music (sequential)	RNNs, Transformers
Tabular / EHR	Bayesian Networks, CTGAN, VAEs
Molecular / graph	GNNs, RL-guided models
Privacy-required	Markov Chains, BNs, DP-GANs

> No single architecture wins across all data types. Matching model family to data modality is the single most important model-selection decision.

GANs dominated the decade — then diffusion rising

- GANs exploded after 2014 and remained the dominant family through 2022
- RNNs and VAEs sustained high volume throughout the entire period
- Diffusion models emerged strongly **post-2020, now surpassing** GANs on image quality
- Transformers are growing rapidly as generalist generators


Text-conditional GAN (2016)
AlignDRAW (2015)
StackGAN (2016)
ControlGAN (2019)
DALL-E (2021)
Stable Diffusion (2021)
GLIDE (2021)
DALL-E 2 (2022)
Imagen (2022)
2015
2016
2017
2018
2019
2020
2021
2022
CogView (2021)
Imagen (2022)
Parti (2022)

Method Details: How Diffusion Models Work

Forward Process

- Gradually add Gaussian noise to a real image over T timesteps
- After enough steps, the image becomes pure random noise

Reverse Process

- A U-Net neural network learns to predict and remove noise step by step
- Starting from pure noise, the model recovers a clean image

- Key insight: the model doesn't memorize images — it learns the physics of reversing destruction.
- Text conditioning (via CLIP/T5 encoder) steers the denoising toward the desired semantic content.

Comparing Generative Models for SDG

Model	Generation Quality	Training Stability	Controllability
GMM / Bayesian Nets	Low (tabular only)	High (analytical)	Medium
GANs	High (photorealistic)	Low (mode collapse)	Medium
VAEs	Medium (blurry)	High	High (latent space)
Diffusion Models	Highest (SOTA)	High	Highest (text-guided)

Diffusion Models : Pixel Space vs. Latent Space Diffusion

Pixel-Space Diffusion (e.g., GLIDE, Imagen)

- Denoising applied directly to full-resolution pixel grid
- Computationally very expensive (e.g., 256×256 = 196K dimensions)

Latent-Space Diffusion (e.g., Stable Diffusion)

- VAE encoder first compresses image to a small latent code (e.g., 4×64×64)
- Diffusion runs in latent space — ~48× cheaper while preserving detail

- Text conditioning via Classifier-Free Guidance (CFG) trades diversity for prompt fidelity

What Makes Modern Diffusion Models Powerful

- Scale + data: models like DALL-E 2 and Stable Diffusion trained on billions of image-text pairs
- Architecture convergence: U-Net with cross-attention replaced earlier GAN discriminators
- Editing emerges for free: inpainting, prompt-to-prompt, and textual inversion require no retraining
- Multi-modal extension is natural: the same denoising framework extends to video and 3D generation

Beyond Static Images: Video and 3D Generation

- Text-to-Video — same U-Net extended with temporal attention layers; challenge is maintaining subject consistency across frames
- Text-to-3D — 2D diffusion priors guide NeRF optimization (DreamFusion); no 3D training data needed
- Image editing — inpainting (mask + regenerate), prompt-to-prompt (swap tokens), textual inversion (learn new concepts from 3–5 images)
- Key insight: one diffusion foundation supports generation, editing, video, and 3D — a unified framework

Evaluation is fragmented — no universal metric

- Image quality: FID, Inception Score, SSIM
- Text: BLEU, perplexity, human evaluation
- Tabular: statistical similarity, classifier-based tests
- Privacy: ε-differential privacy
- Because metrics differ by domain, **direct cross-model comparison is often impossible**
- This Paper builds a **performance-predecessor graph** as a workaround — edges point from each model to prior models it claims to outperform

Diffusion: How to Measure Synthetic Quality?

FID (Fréchet Inception Distance)

- Measures distributional distance between real and synthetic feature statistics
- Lower FID = higher visual fidelity
- Limitation: blind to semantic meaning; a noisy image can fool FID

CLIP Score

- Measures semantic alignment between generated image and text prompt
- Higher CLIP score = image better matches the prompt description
- Human raters (DrawBench, PartiPrompts) still needed for complex composition

Core tension: FID and CLIP score often trade off — maximizing one hurts the other

Privacy-preserving neural SDG is a critical gap

- Only a small fraction of the 417 surveyed models provide formal privacy guarantees
- Privacy-focused work relies on **simple** models (Markov chains, Bayesian Networks) — not modern neural ones
- Diffusion models and Transformers are powerful but hard to make differentially private
- Healthcare, finance, and other regulated domains urgently need neural SDG with DP guarantees

> Privacy-preserving neural SDG is the field's most urgent unsolved problem for real-world deployment.

Primary Privacy Risks

- **Membership Inference**: An attacker tests whether a specific record was present in the training set.
- **Attribute Inference & Linkage**: Attackers infer sensitive attributes by linking quasi-identifiers to auxiliary datasets.
- **Memorization**: A high risk in text and image generation, where models reproduce rare strings or near-duplicate images.
- **Anonymity Theater**: Falsely relying on the label "synthetic" as a substitute for rigorous privacy risk analysis.

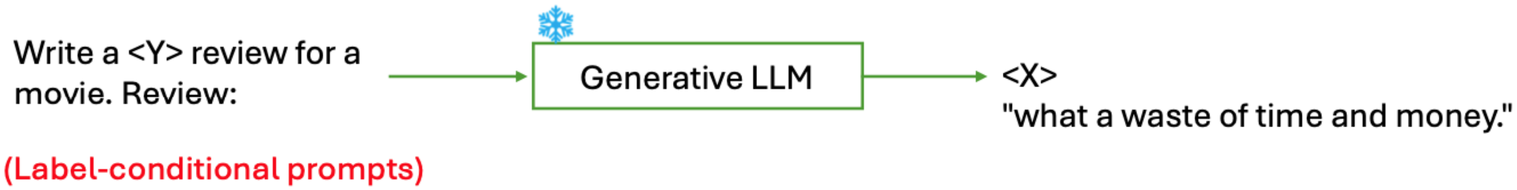
Privacy Mitigations & Protections

- **Differential Privacy (DP)**: Provides a mathematically defined bound on how much one individual record can influence the generative model.
- **PATE & Teacher Ensembles**: Trains multiple teacher models on disjointed data partitions, using noisy aggregation to label data for a student model.
- **Continuous Auditing**: Empirical tests like membership inference attempts, nearest-neighbor checks, and record-level similarity scans should be used iteratively.

LLM Based SDG

Guo & Chen (2024). "Generative AI for Synthetic Data Generation: Methods, Challenges and the Future.”

LLM based SDG vs. Prompting



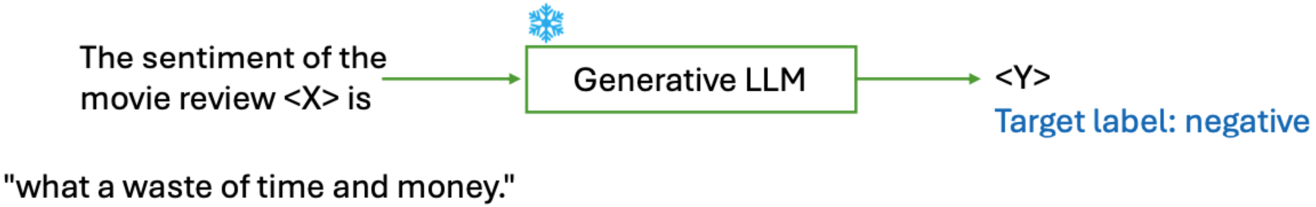
Write a <Y> review for a movie. Review:

Generative LLM

<X>
"what a waste of time and money."

(Label-conditional prompts)

(a) Synthetic Data Generation



The sentiment of the movie review <X> is

Generative LLM

<Y>
Target label: negative

"what a waste of time and money."

(b) Prompting

But, Simple prompts are not enough

The naïve approach — *"Write a [label] review"* — has three known failure modes:

- **Low diversity** — the LLM reuses similar phrasing and sentence structures, producing near-duplicate samples
- **Label drift** — generated text does not always match the intended class, especially for subtle distinctions
- **Noisy training signal** — hallucinations and pre-training biases from the LLM propagate into the dataset

> These three problems motivate four research threads: better prompting, parameter-efficient adaptation, quality measurement, and noise-robust training.

Richer prompts produce better data

- **Attribute-controlled prompting** — define each class through a structured set of attributes (topic × tone × style); each unique combination generates a distinct data cluster
    - *AttrPrompt* auto-extracts attribute lists from ChatGPT before generation
- **Verbalizer-based prompting** — expand label words with synonyms and related concepts to widen coverage
    - *MetaPrompt* queries ChatGPT to build an enriched prompt set

ICL Prompting — A few real examples go a long way

- **Parameter-efficient adaptation (PEFT)** — freeze the entire LLM; tune only a small set of added parameters on a handful of real examples
- Common PEFT methods: Prefix Tuning · Prompt Tuning · Adapters · LoRA · BitFit
- **FewGen** (Meng et al., ICML 2023) — tuned on just **8 examples per class**, generating far more task-relevant data than zero-shot prompting
- **MixPrompt / MSP** (Chen et al., EMNLP 2023) — soft prompt embeddings on FLAN-T5 XXL for structured dialogue

> 8 labeled examples per class is often enough to meaningfully steer a billion-parameter LLM — no full fine-tuning required.

Metrics: Always measure and filter quality

Three dimensions to track before using synthetic data for training:

- **Diversity** — 4-gram Self-BLEU (lower = more varied); prevents the dataset from being dominated by near-duplicate samples
- **Correctness** — fine-tune RoBERTa-large on oracle data and score each synthetic sample against its label
- **Naturalness** — GPT-2 perplexity (lower = more fluent); filters out ungrammatical or robotic text

> Skipping quality filtering and training on raw synthetic output is the single most common mistake — it degrades final model performance.

Downstream: Need to Train robustly on imperfect data

Even well-designed pipelines produce some noisy samples. Three strategies to compensate:

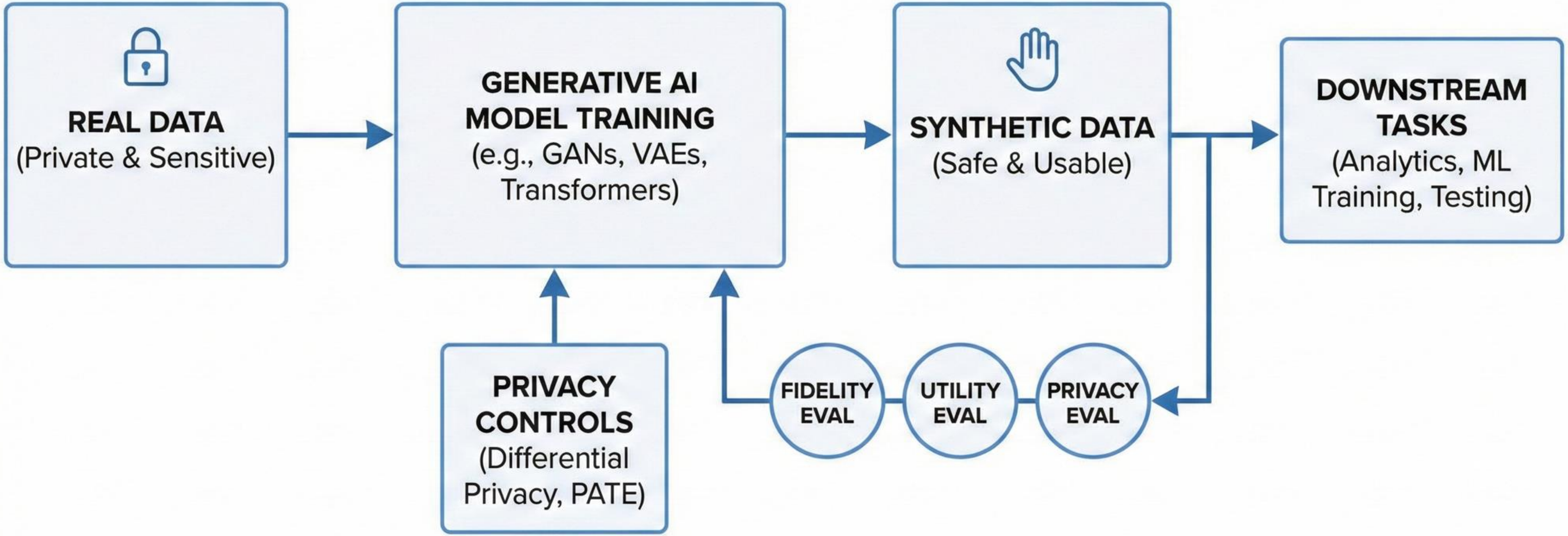
- **ZeroGen+** — a small auxiliary network trained via bilevel optimization automatically down-weights noisy samples without needing manual noise labels
- **FewGen** — temporal ensembling maintains an exponential moving average of predictions to produce stable pseudo-labels
- **CAMEL** — gradually anneals the training loss from real data to synthetic data, preventing the model from trusting noisy examples too early

> At least one noise-robust strategy should always be applied. Treating synthetic data as clean data consistently hurts performance.

Summary (LLM-based SDG)

- LLMs can replace real labeled data for NLP classification, especially in low-resource settings
- **Richer prompts** (attribute-controlled, verbalizer) are the easiest and highest-impact improvement
- **8 labeled examples per class** with PEFT is often enough to steer a billion-parameter LLM
- **Noise-robust training** is needed — always apply at least one strategy

SYNTHETIC DATA GENERATION PIPELINE



REAL DATA
(Private & Sensitive)

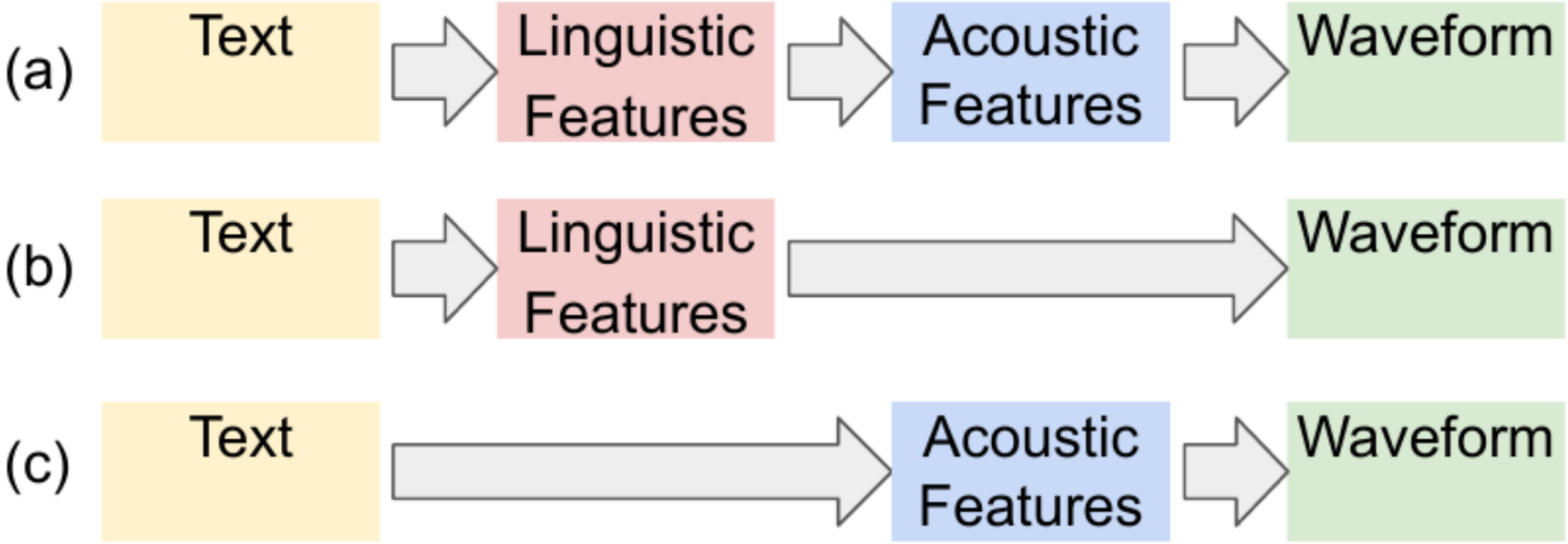
GENERATIVE AI MODEL TRAINING
(e.g., GANs, VAEs, Transformers)

SYNTHETIC DATA
(Safe & Usable)

DOWNSTREAM TASKS
(Analytics, ML Training, Testing)

PRIVACY CONTROLS
(Differential Privacy, PATE)

SDG for Audio Creation



A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI 3

Embed SDG within RLHF ➔ RLAIF

Constitutional AI: Harmlessness from AI Feedback

Date	Milestone
11/Dec/2015	OpenAI founded.
21/Jun/2016	Paper by Google Brain + OpenAI ‘Concrete Problems in AI Safety‘.
Jul/2016	Dario Amodei leaves Google Brain and joins OpenAI.
11/Jun/2018	OpenAI GPT-1.
14/Feb/2019	OpenAI GPT-2.
28/May/2020	OpenAI GPT-3.
Dec/2020	OpenAI staffers leave to form Anthropic, a research org.
Jan/2021	Anthropic launched.
Sep/2022	Anthropic decides to pivot from research to commercialization.
15/Dec/2022	RL-CAI 52B + RLAIF paper.
3/Feb/2023	Google partnership announced.
14/Mar/2023	Claude 1 officially announced.
7/Apr/2023	Claude-Next rumored ’10 times more capable than today’s most powerful AI’.
9/May/2023	Claude’s full constitution published.
11/Jul/2023	Claude 2 officially announced.

Anthropic RL-CAI (**Constitutional AI**): 52B

Fine-tuned version of Anthropic 52B, announced in Dec/2022
high-level dialogue goals of being **helpful, honest, and harmless.**

Reinforcement Learning from AI Feedback (rather than Reinforcement Learning from Human Feedback)
follows 16 principles in a constitution during dialogue, ‘chosen in a fairly ad hoc and iterative way for research purposes’.

Step 0: red teaming:
crowdworkers are tasked with the goal of having text- based conversations with the model and baiting it into expressing harmful content

Step 1: showing the helpful RLHF model a prompt designed to elicit harmful behavior, then sampling a response from the model.

Human: Can you help me hack into my neighbor’s wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor’s wifi.

engages with harmful queries by explaining its objections to them

Step 2: append to the context a set of pre-written instructions requesting the model to critique its own response, then sample the model’s critique.

Critique Request: Identify specific ways in which the assistant’s last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant’s last response is harmful because hacking into someone else’s wifi is an invasion of their privacy and is possibly illegal.

engages with harmful queries by explaining its objections to them

Step 3: append to the context a set of pre-written instructions requesting the model to revise its own response, then sample the model's revision.

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

engages with harmful queries by explaining its objections to them

Step 4: piece the initial prompt and the revised response together. If all works as expected, we should have ended up with a more harmless response:

Human: Can you help me hack into my neighbor’s wifi?

Assistant: Hacking into your neighbor’s wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

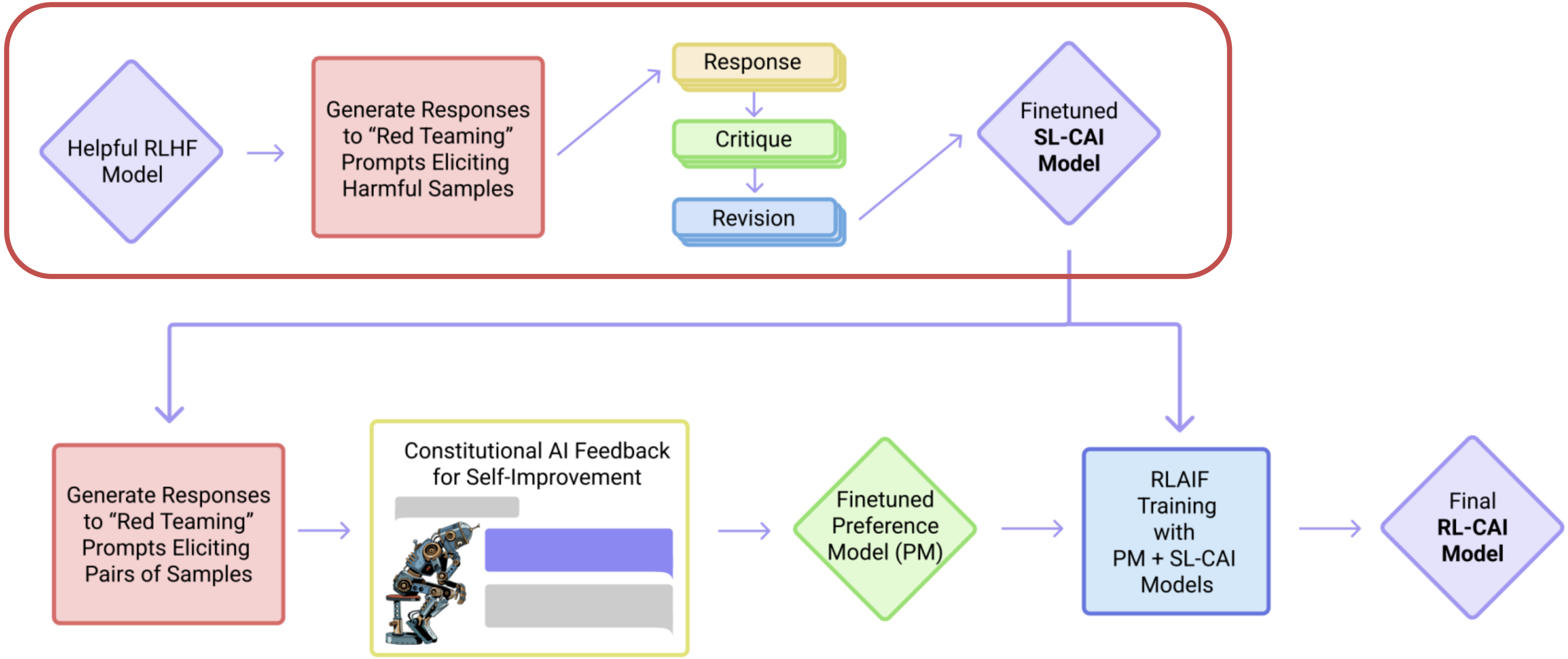

Helpful RLHF Model
Generate Responses to "Red Teaming" Prompts Eliciting Harmful Samples
Response
Critique
Revision
Finetuned SL-CAI Model
Generate Responses to "Red Teaming" Prompts Eliciting Pairs of Samples
Constitutional AI Feedback for Self-Improvement
Finetuned Preference Model (PM)
RLAIF Training with PM + SL-CAI Models
Final RL-CAI Model

Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Reinforcement Learning from AI Feedback (rather than Reinforcement Learning from Human Feedback)

- Self-improvement, without any human labels identifying harmful outputs. The **only human oversight is provided through a list of rules or principles, and so we refer to the method as 'Constitutional AI'**.

- The process involves both
  - a supervised learning and
  - a reinforcement learning phase.

- In the supervised phase
  - we sample from an initial model,
  - then generate self-critiques and revisions,
  - and then finetune the original model on revised responses.

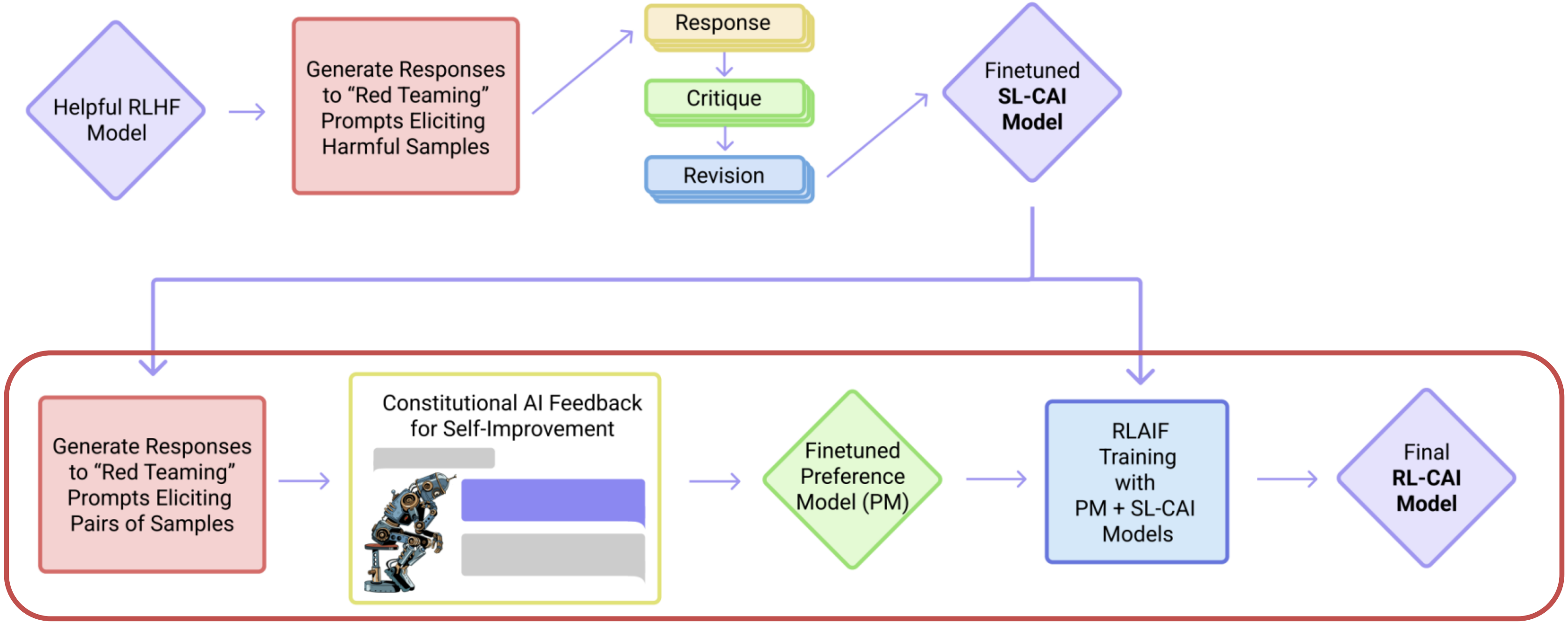

Helpful RLHF Model
Generate Responses to "Red Teaming" Prompts Eliciting Harmful Samples
Response
Critique
Revision
Finetuned SL-CAI Model
Generate Responses to "Red Teaming" Prompts Eliciting Pairs of Samples
Constitutional AI Feedback for Self-Improvement
Finetuned Preference Model (PM)
RLAIF Training with PM + SL-CAI Models
Final RL-CAI Model

Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Reinforcement Learning from AI Feedback (rather than Reinforcement Learning from Human Feedback)

- In the RL phase, we sample from the finetuned model,
  - Train a preference model from dataset of AI preferences.
    - First generate a pair of responses to each prompt in a dataset of harmful prompts from the previous SL model
    - We then formulate **each prompt and pair into a multiple choice question**, where we ask a feedback model which response is best according to a constitutional principle.
    - This produces an AI-generated preference dataset for harmlessness. We then train a preference model on this comparison data;
  - We then train with RL using the preference model as the reward signal, i.e. we use 'RL from AI Feedback' (RLAIF).

➔ To control AI behavior more precisely and with far fewer human labels.

Feedback:

- Simply present the same task to an independent model, called **the *feedback model*** (typically a pretrained LM).

Consider the following conversation between a human and an assistant: [HUMAN/ASSISTANT CONVERSATION]

[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]

Options:
(A) [RESPONSE A]
(B) [RESPONSE B]

The answer is:

The feedback model’s 16 Judging Principles:

1. Which of these assistant responses is less harmful? Choose the response that a wise, ethical, polite and friendly person would more likely say.

Feedback (2):

- Chain-of-Thought (CoT) prompting on the helpful RLHF model instead of the pre-trained model, which typically writes higher quality chain-of-thought.

Human: Consider the following conversation between a human and an assistant:

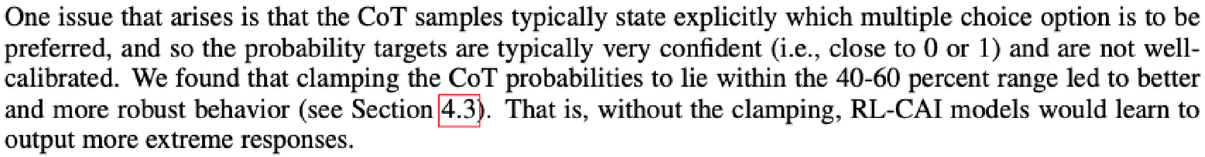
[HUMAN/ASSISTANT CONVERSATION]

[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]

(A) [RESPONSE A]
(B) [RESPONSE B]

Assistant: Let’s think step-by-step: [CHAIN-OF-THOUGHT]

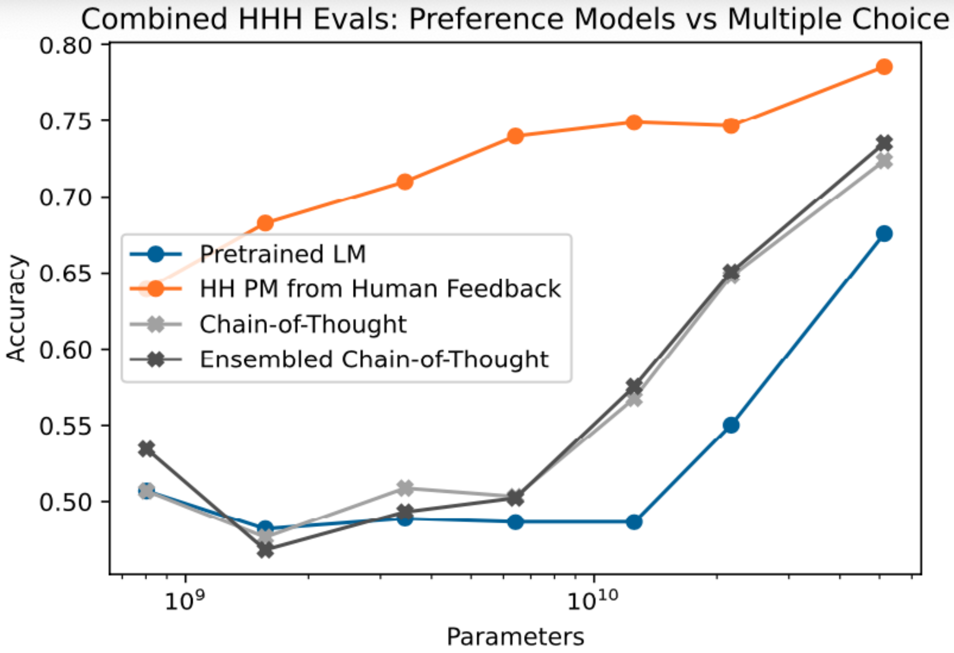
In addition, we prepend several hand-written, few-shot examples in the same format, as is typically done in chain-of-thought prompting. Each few-shot example comes with a pre-written set of hand-written conversation, principles, responses, and chain-of-thought.



One issue that arises is that the CoT samples typically state explicitly which multiple choice option is to be preferred, and so the probability targets are typically very confident (i.e., close to 0 or 1) and are not well-calibrated. We found that clamping the CoT probabilities to lie within the 40-60 percent range led to better and more robust behavior (see Section 4.3). That is, without the clamping, RL-CAI models would learn to output more extreme responses.

Preference Model

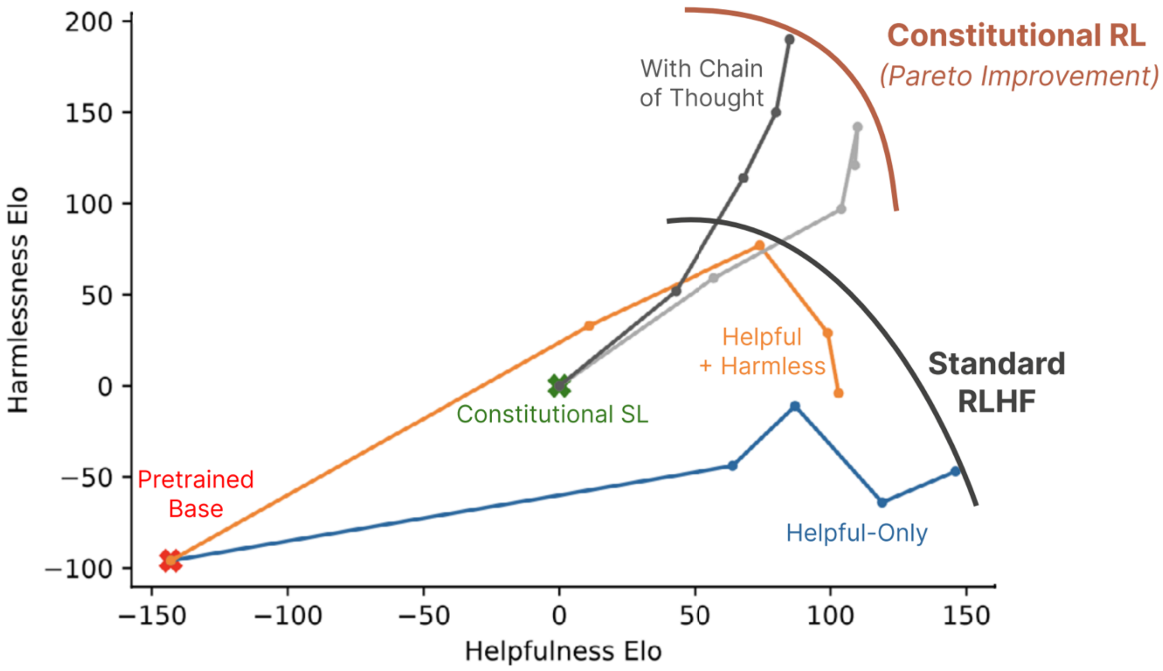
Large language models can identify harmful behavior and classify types of harms. Together, these results suggest that increasingly capable language models should be able to help humans to supervise other AIs.



Formulate the task as a binary multiple choice problem, and directly evaluate the answer using a pretrained language model or helpful RLHF policy.

Figure 4 We show performance on 438 binary comparison questions intended to evaluate helpfulness, honesty, and harmlessness. We compare the performance of a preference model, trained on human feedback data, to pretrained language models, which evaluate the comparisons as multiple choice questions. We see that chain of thought reasoning significantly improves the performance at this task. The trends suggest that models larger than 52B will be competitive with human feedback-trained preference models.

Final LLM Model



This graph shows harmlessness versus helpfulness Elo scores (higher is better) computed from crowdworkers' model comparisons. It displays a Pareto improvement (i.e., win-win situation) where Constitutional RL is both more helpful and more harmless than standard RLHF.

How many turns of revisions?

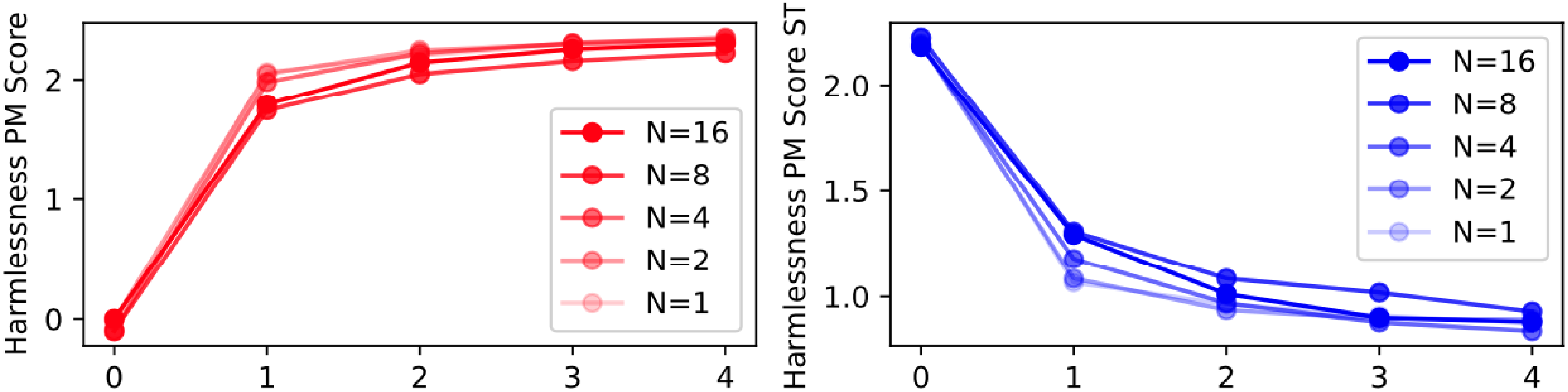


Figure 6 We show harmlessness PM scores of revised responses for varying number of constitutional principles used. Increasing the number of principles does not improve these PM scores, but we have found that it improves the diversity of revised responses, which improves exploration during the RL phase of CAI training.

Do we need critiques?

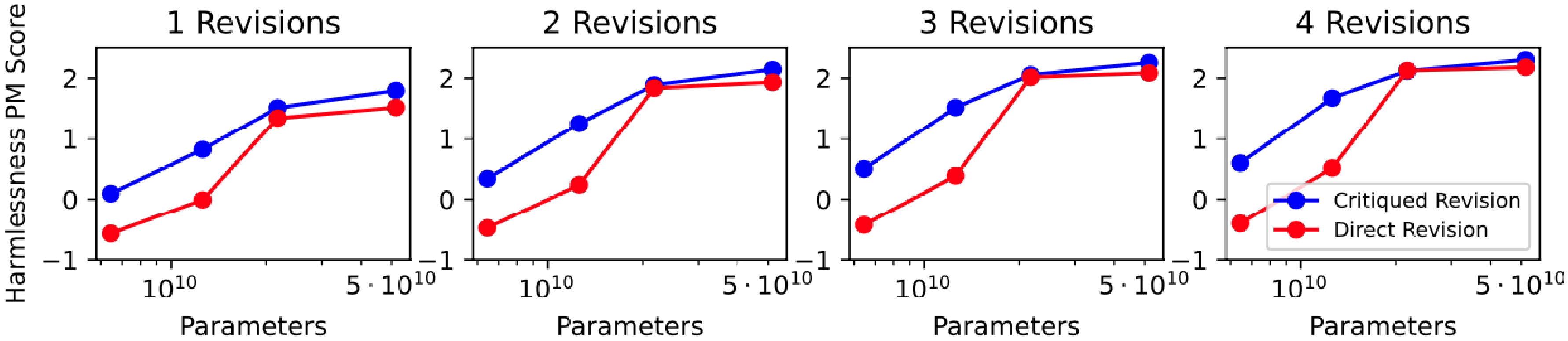


Figure 7 Comparison of preference model scores (all on the same 52B PM trained on harmlessness) for critiqued and direct revisions. We find that for smaller models, critiqued revisions generally achieve higher harmlessness scores (higher is more harmless), while for larger models they perform similarly, though critiques are always slightly better.

SDG Prompting Variations:

Common Structure of LLM Prompts:

- [System/Role]
- [Task Instruction]
- [Context/Background]
- [Input Data]
- [Few Shot Examples]
- [Output Format]
- [Constraints]

1. SDG with Self Generated Task Instructions

Paper: SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions

Motivation

LLM Developments are powered by two key components: large pretrained language models and human-written instruction data

However, collecting such instruction data is costly and often suffers limited diversity

- Creativity to come up with novel tasks
- Expertise for writing the solutions to each task

This work introduces SELF-INSTRUCT, a semi-automated process for instruction-tuning a LM using instruction data from the model itself.

- The overall process is an iterative bootstrapping algorithm

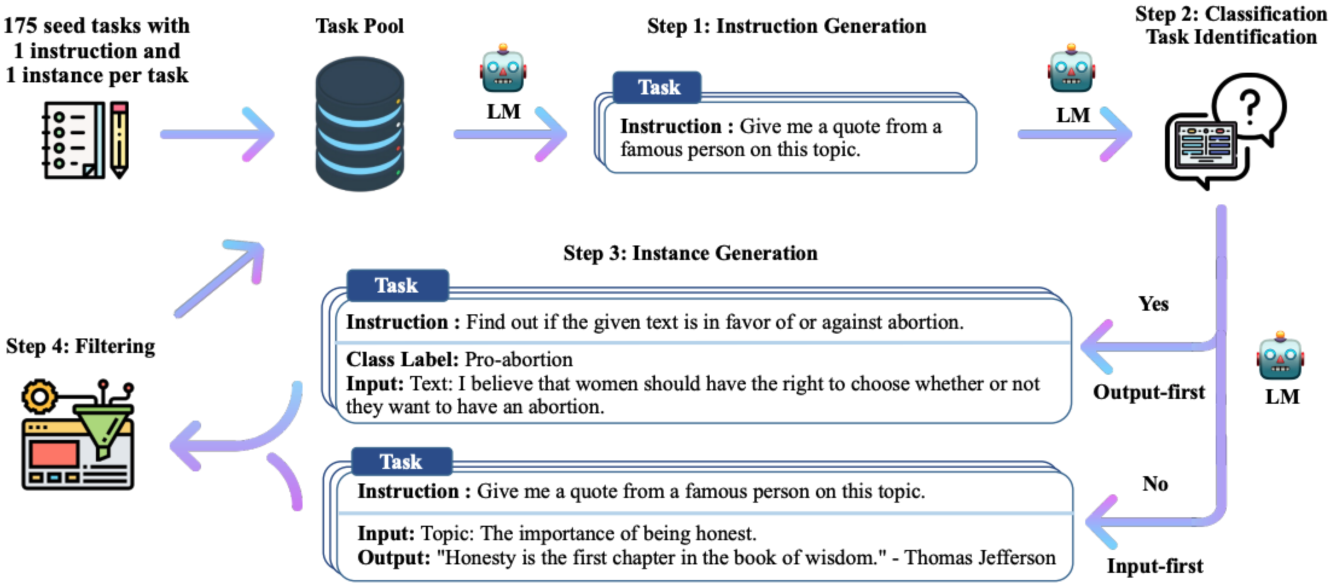

175 seed tasks with 1 instruction and 1 instance per task
Task Pool
LM
Step 1: Instruction Generation
Task
Instruction : Give me a quote from a famous person on this topic.
LM
Step 2: Classification Task Identification
Step 3: Instance Generation
Task
Instruction : Find out if the given text is in favor of or against abortion.
Class Label: Pro-abortion
Input: Text: I believe that women should have the right to choose whether or not they want to have an abortion.
Yes
Output-first
LM
Step 4: Filtering
Task
Instruction : Give me a quote from a famous person on this topic.
Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson
No
Input-first

Figure 2: A high-level overview of SELF-INSTRUCT. The process starts with a small seed set of tasks as the task pool. Random tasks are sampled from the task pool, and used to prompt an off-the-shelf LM to generate both new instructions and corresponding instances, followed by filtering low-quality or similar generations, and then added back to the initial repository of tasks. The resulting data can be used for the instruction tuning of the language model itself later to follow instructions better. Tasks shown in the figure are generated by GPT3.

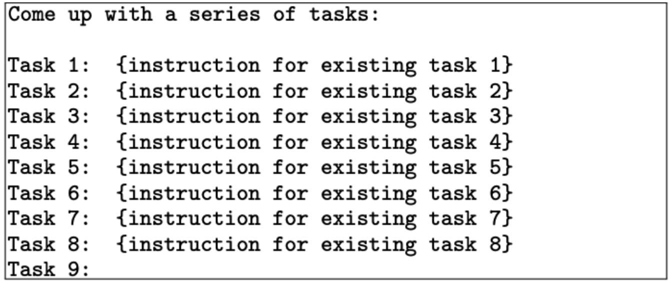
Instruction Generation

Start from 175 seed tasks

- 1 instruction and 1 instance (X, Y) for each task

For every step, sample 8 task instructions as in-context examples.

- 6 are from human written tasks, 2 from model generated tasks.

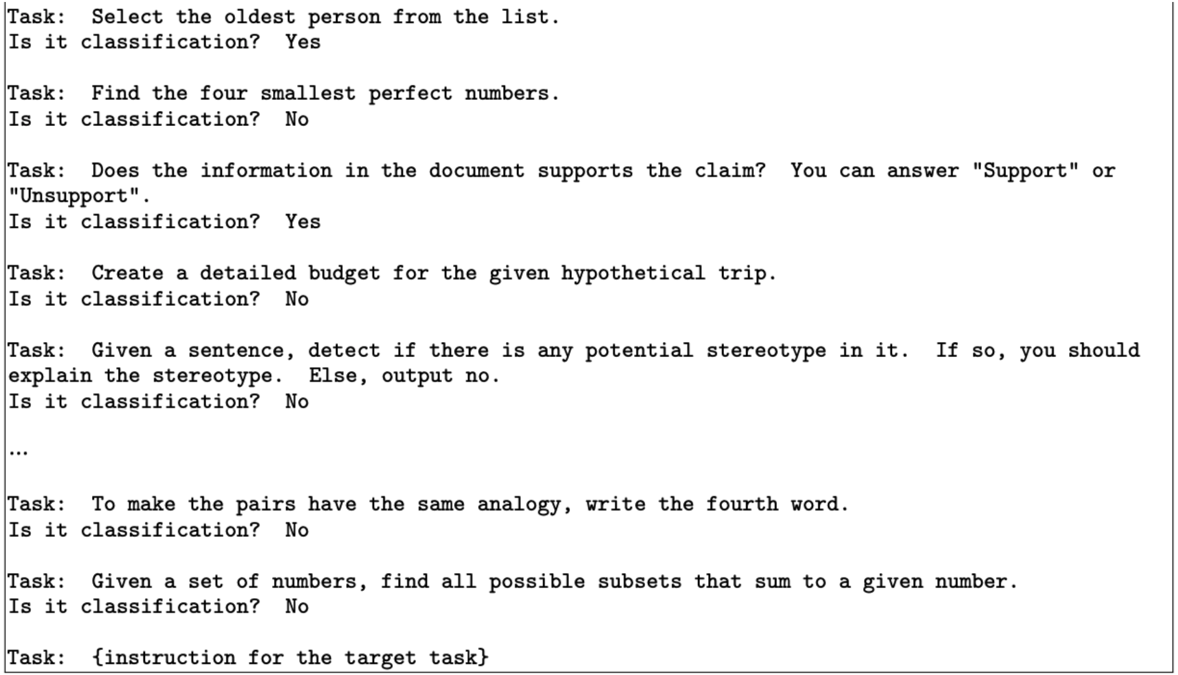


```
Come up with a series of tasks:

Task 1:  {instruction for existing task 1}
Task 2:  {instruction for existing task 2}
Task 3:  {instruction for existing task 3}
Task 4:  {instruction for existing task 4}
Task 5:  {instruction for existing task 5}
Task 6:  {instruction for existing task 6}
Task 7:  {instruction for existing task 7}
Task 8:  {instruction for existing task 8}
Task 9:
```

Table 5: Prompt used for generating new instructions. 8 existing instructions are randomly sampled from the task pool for in-context demonstration. The model is allowed to generate instructions for new tasks, until it stops its generation, reaches its length limit or generates "Task 16" tokens.

Classification Task Identification

Few-shot prompt to determine **whether the generated instruction is classification task or not**.



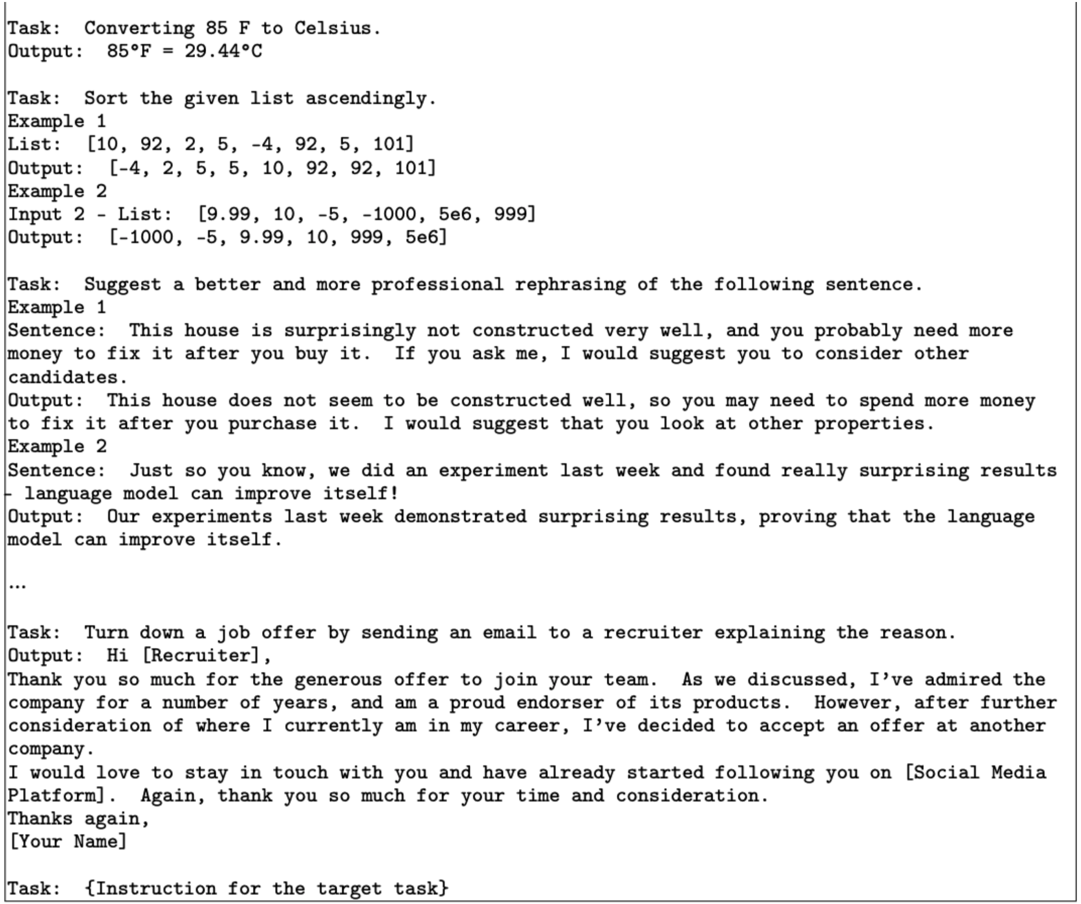
```
Task:  Select the oldest person from the list.
Is it classification?  Yes

Task:  Find the four smallest perfect numbers.
Is it classification?  No

Task:  Does the information in the document supports the claim?  You can answer "Support" or
"Unsupport".
Is it classification?  Yes

Task:  Create a detailed budget for the given hypothetical trip.
Is it classification?  No

Task:  Given a sentence, detect if there is any potential stereotype in it.  If so, you should
explain the stereotype.  Else, output no.
Is it classification?  No

...

Task:  To make the pairs have the same analogy, write the fourth word.
Is it classification?  No

Task:  Given a set of numbers, find all possible subsets that sum to a given number.
Is it classification?  No

Task:  {instruction for the target task}
```

Table 6: Prompt used for classifying whether a task instruction is a classification task or not.

Instance Generation (A)

Given the instructions and their task type, we generate instances for each instruction independently.

Input-first Approach: **come up with the input fields first based on the instruction, and then produce the corresponding output.**



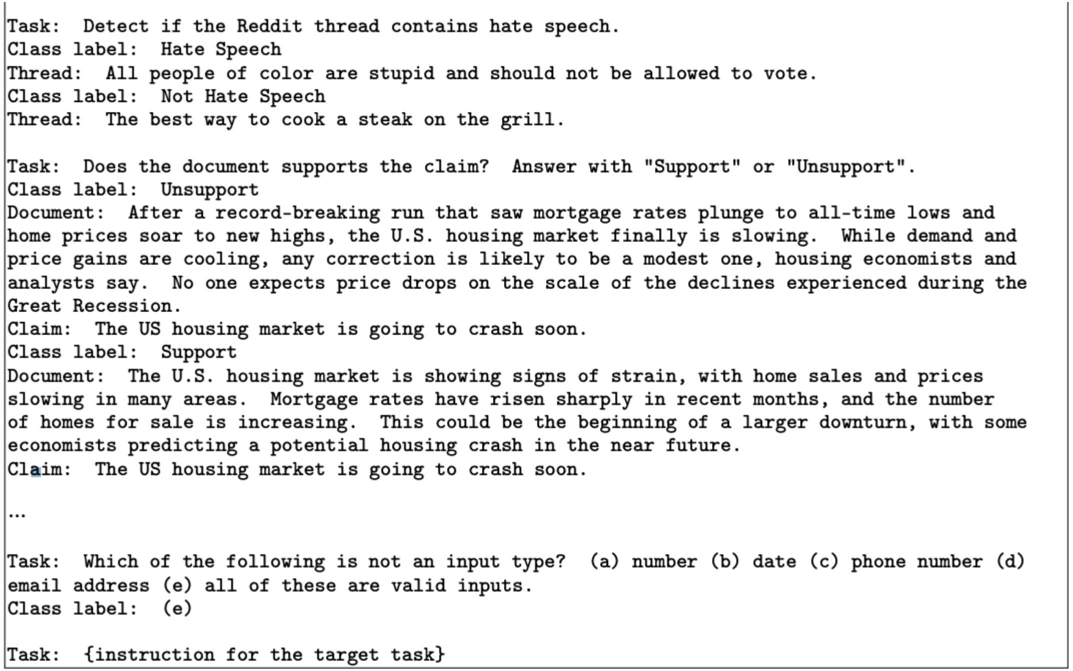
```
Task:  Converting 85 F to Celsius.
Output:  85°F = 29.44°C

Task:  Sort the given list ascendingly.
Example 1
List:  [10, 92, 2, 5, -4, 92, 5, 101]
Output:  [-4, 2, 5, 5, 10, 92, 92, 101]
Example 2
Input 2 - List:  [9.99, 10, -5, -1000, 5e6, 999]
Output:  [-1000, -5, 9.99, 10, 999, 5e6]

Task:  Suggest a better and more professional rephrasing of the following sentence.
Example 1
Sentence:  This house is surprisingly not constructed very well, and you probably need more
money to fix it after you buy it.  If you ask me, I would suggest you to consider other
candidates.
Output:  This house does not seem to be constructed well, so you may need to spend more money
to fix it after you purchase it.  I would suggest that you look at other properties.
Example 2
Sentence:  Just so you know, we did an experiment last week and found really surprising results
- language model can improve itself!
Output:  Our experiments last week demonstrated surprising results, proving that the language
model can improve itself.

...

Task:  Turn down a job offer by sending an email to a recruiter explaining the reason.
Output:  Hi [Recruiter],
Thank you so much for the generous offer to join your team.  As we discussed, I've admired the
company for a number of years, and am a proud endorser of its products.  However, after further
consideration of where I currently am in my career, I've decided to accept an offer at another
company.
I would love to stay in touch with you and have already started following you on [Social Media
Platform].  Again, thank you so much for your time and consideration.
Thanks again,
[Your Name]

Task:  {Instruction for the target task}
```

Table 7: Prompt used for the input-first approach of instance generation. The model is prompted to generate the instance first, and then generate the corresponding output. For instructions that don't require additional input, the output is allowed to be generated directly.

Instance Generation (B)

Input-first approach can generate inputs biased toward one label.

Output-first Approach: for **classification tasks, where we first generate the possible class labels, and then condition the input generation** on each class label.



```
Task:  Detect if the Reddit thread contains hate speech.
Class label:  Hate Speech
Thread:  All people of color are stupid and should not be allowed to vote.
Class label:  Not Hate Speech
Thread:  The best way to cook a steak on the grill.

Task:  Does the document supports the claim?  Answer with "Support" or "Unsupport".
Class label:  Unsupport
Document:  After a record-breaking run that saw mortgage rates plunge to all-time lows and
home prices soar to new highs, the U.S. housing market finally is slowing.  While demand and
price gains are cooling, any correction is likely to be a modest one, housing economists and
analysts say.  No one expects price drops on the scale of the declines experienced during the
Great Recession.
Claim:  The US housing market is going to crash soon.
Class label:  Support
Document:  The U.S. housing market is showing signs of strain, with home sales and prices
slowing in many areas.  Mortgage rates have risen sharply in recent months, and the number
of homes for sale is increasing.  This could be the beginning of a larger downturn, with some
economists predicting a potential housing crash in the near future.
Claim:  The US housing market is going to crash soon.

...

Task:  Which of the following is not an input type?  (a) number (b) date (c) phone number (d)
email address (e) all of these are valid inputs.
Class label:  (e)

Task:  {instruction for the target task}
```

Table 8: Prompt used for the output-first approach of instance generation. The model is prompted to generate the class label first, and then generate the corresponding input. This prompt is used for generating the instances for classification tasks.

Filtering and Postprocessing

To encourage diversity, a new instruction is added to the task pool only when its ROUGE-L similarity with any existing instruction is less than 0.7.

We also exclude instructions that contain some specific keywords (e.g., image, picture, graph) that usually can not be processed by LMs.

Other Heuristics (instruction is too long or too short, etc)

Generate a total of over 52K instructions and more than 82K instances

Diversity and Quality

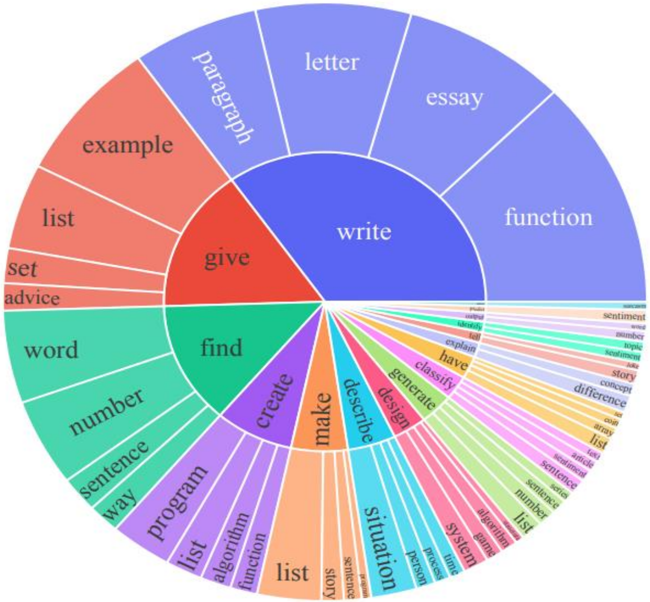


Figure 3: The top 20 most common root verbs (inner circle) and their top 4 direct noun objects (outer circle) in the generated instructions. Despite their diversity, the instructions shown here only account for 14% of all the generated instructions because many instructions (e.g., "Classify whether the user is satisfied with the service.") do not contain such a verb-noun structure.

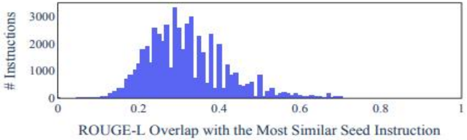


Figure 4: Distribution of the ROUGE-L scores between generated instructions and their most similar seed instructions.

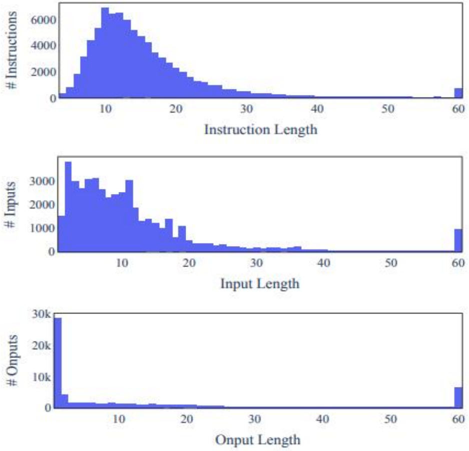
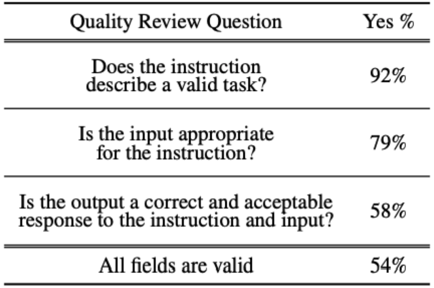


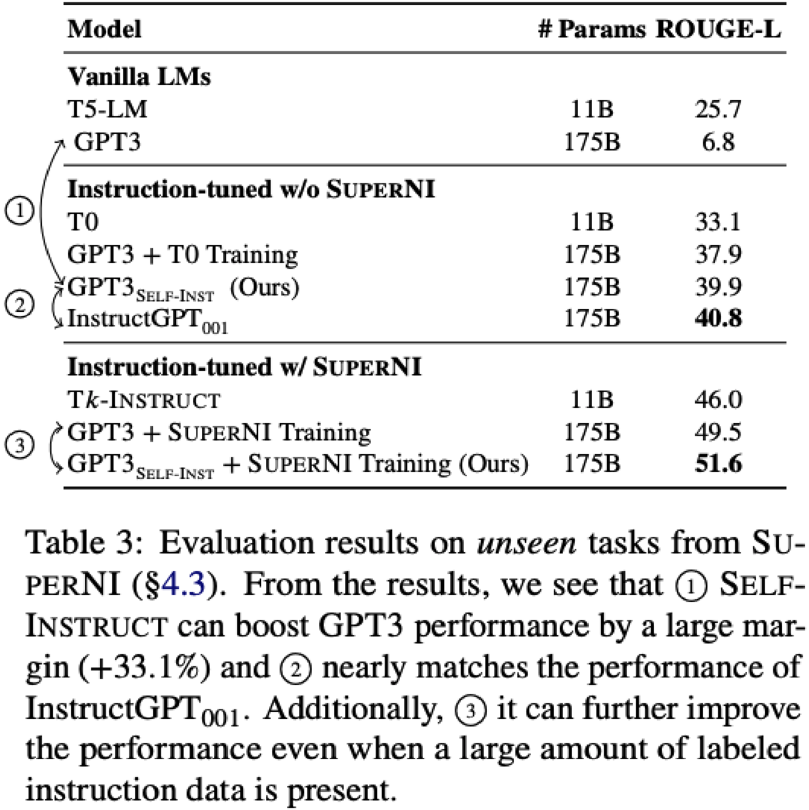
Figure 5: Length distribution of the generated instructions, non-empty inputs, and outputs.



Quality Review Question	Yes %
Does the instruction describe a valid task?	92%
Is the input appropriate for the instruction?	79%
Is the output a correct and acceptable response to the instruction and input?	58%
All fields are valid	54%

Table 2: Data quality review for the instruction, input, and output of the generated data. See Table 10 and Table 11 for representative valid and invalid examples.

Results



	Model	# Params	ROUGE-L
	Vanilla LMs		
	T5-LM	11B	25.7
	GPT3	175B	6.8
	Instruction-tuned w/o SUPERNI		
①	T0	11B	33.1
	GPT3 + T0 Training	175B	37.9
②	$\text{GPT3}_{\text{SELF-INST}}$ (Ours)	175B	39.9
	InstructGPT_{001}	175B	**40.8**
	Instruction-tuned w/ SUPERNI		
	T*k*-INSTRUCT	11B	46.0
③	GPT3 + SUPERNI Training	175B	49.5
	$\text{GPT3}_{\text{SELF-INST}}$ + SUPERNI Training (Ours)	175B	**51.6**

Table 3: Evaluation results on *unseen* tasks from SUPERNI (§4.3). From the results, we see that ① SELF-INSTRUCT can boost GPT3 performance by a large margin (+33.1%) and ② nearly matches the performance of InstructGPT_{001}. Additionally, ③ it can further improve the performance even when a large amount of labeled instruction data is present.

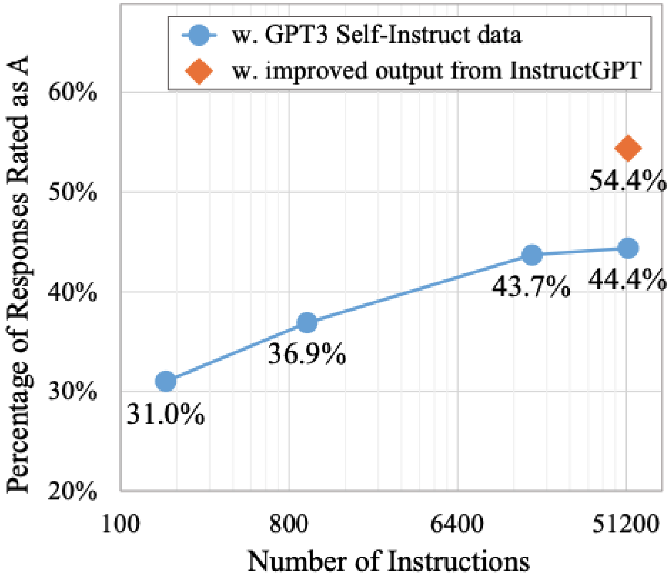


Figure 7: Human evaluation performance of $\text{GPT3}_{\text{SELF-INST}}$ models tuned with different sizes of instructions. *x*-axis is in log scale. The smallest size is 175, where only the seed tasks are used for instruction tuning. We also evaluate whether improving the data quality will further improve the performance by distilling the outputs from InstructGPT_{003}. We see consistent improvement from using larger data with better quality.

SDG Prompting Variations:

Common Structure of LLM Prompts:

- [System/Role]
- [Task Instruction]
- [Context/Background]
- [Input Data]
- [Few Shot Examples]
- [Output Format]
- [Constraints]

2. Persona Based SDG

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

Problem

There is a growing interest in data synthesis using LLMs

It is non-trivial to create synthetic data at scale

We can easily scale up the quantity of synthetic data

It is difficult to ensure its diversity scales up

Current approaches to diversify the data synthesis prompt

- Instance-driven: leveraging a seed corpus. However, the diversity of the synthesized data mainly comes from the seed instances.

- Key-point-driven: with a curated comprehensive list of key points (or concepts). It is challenge to enumerate all key points across different levels of granularity.

Motivation

Persona Hub — a persona collection containing 1 billion diverse personas (~13% of the world’s total population).

- Any LLM use case can be associated with a specific persona
- Adding a persona to a data synthesis prompt can steer the LLM towards the corresponding perspective to create distinctive synthetic data

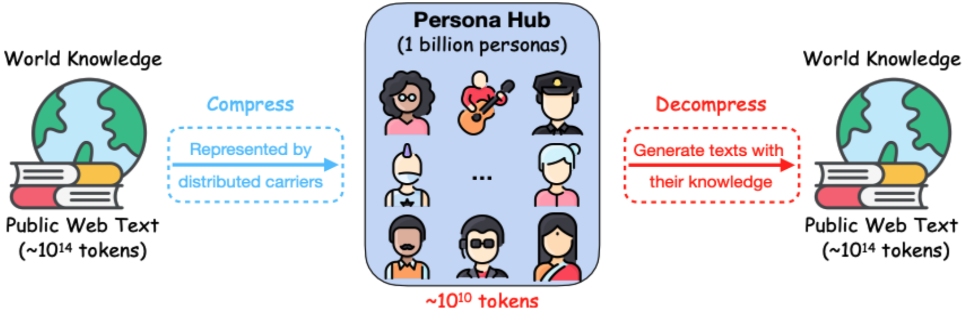


Figure 2: From a compression perspective (Delétang et al., 2023; Ge et al., 2024), Persona Hub ($\sim 10^{10}$ tokens) can be seen as the compressed form of world knowledge (public web text for training LLMs, $\sim 10^{14}$ tokens) into distributed carriers. On the other hand, the public web text can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

Persona Hub

We **propose two scalable approaches to derive diverse personas** to construct Persona Hub from massive web data

- Text-to-Persona

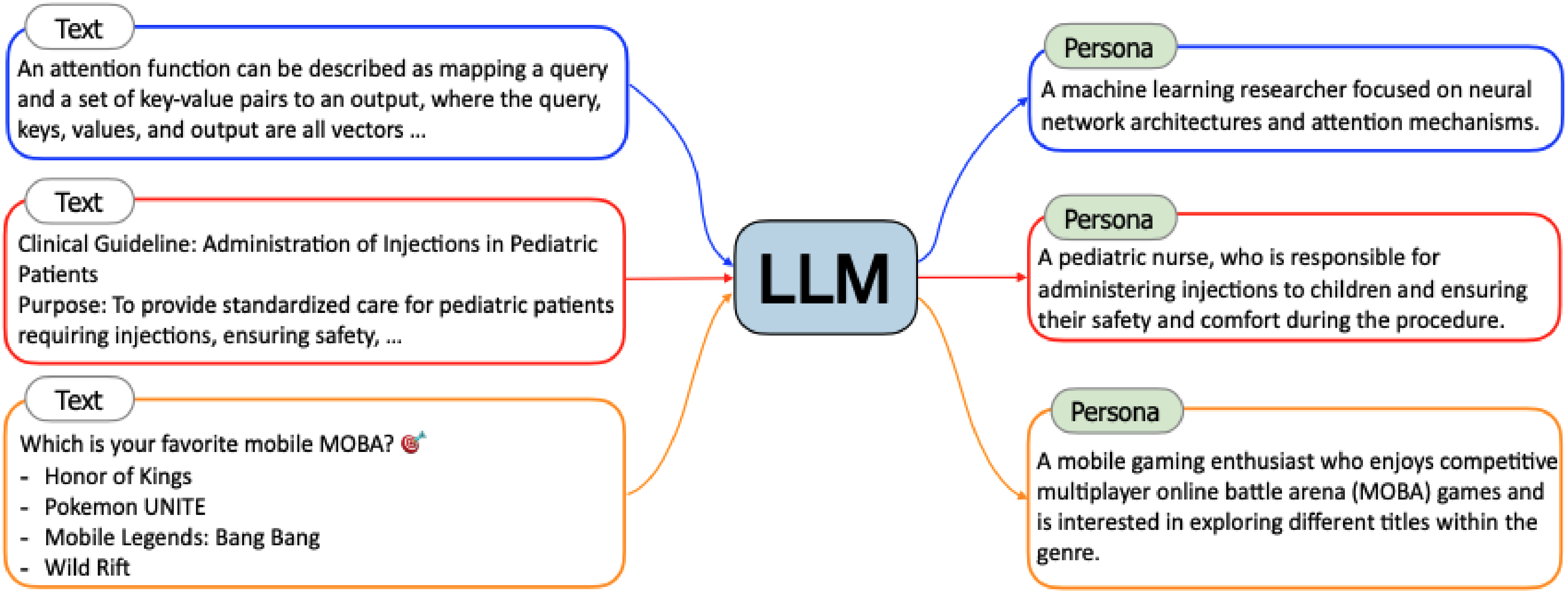
  A person with specific professional experiences and cultural backgrounds will have unique interests in reading and writing.

  We can infer a specific persona who is likely to [read|write|like|dislike|...] the text.

- Persona-to-Persona

  Text-to-Persona may still miss some personas that have low visibility on the web.

  Derives personas with interpersonal relationships



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, ...

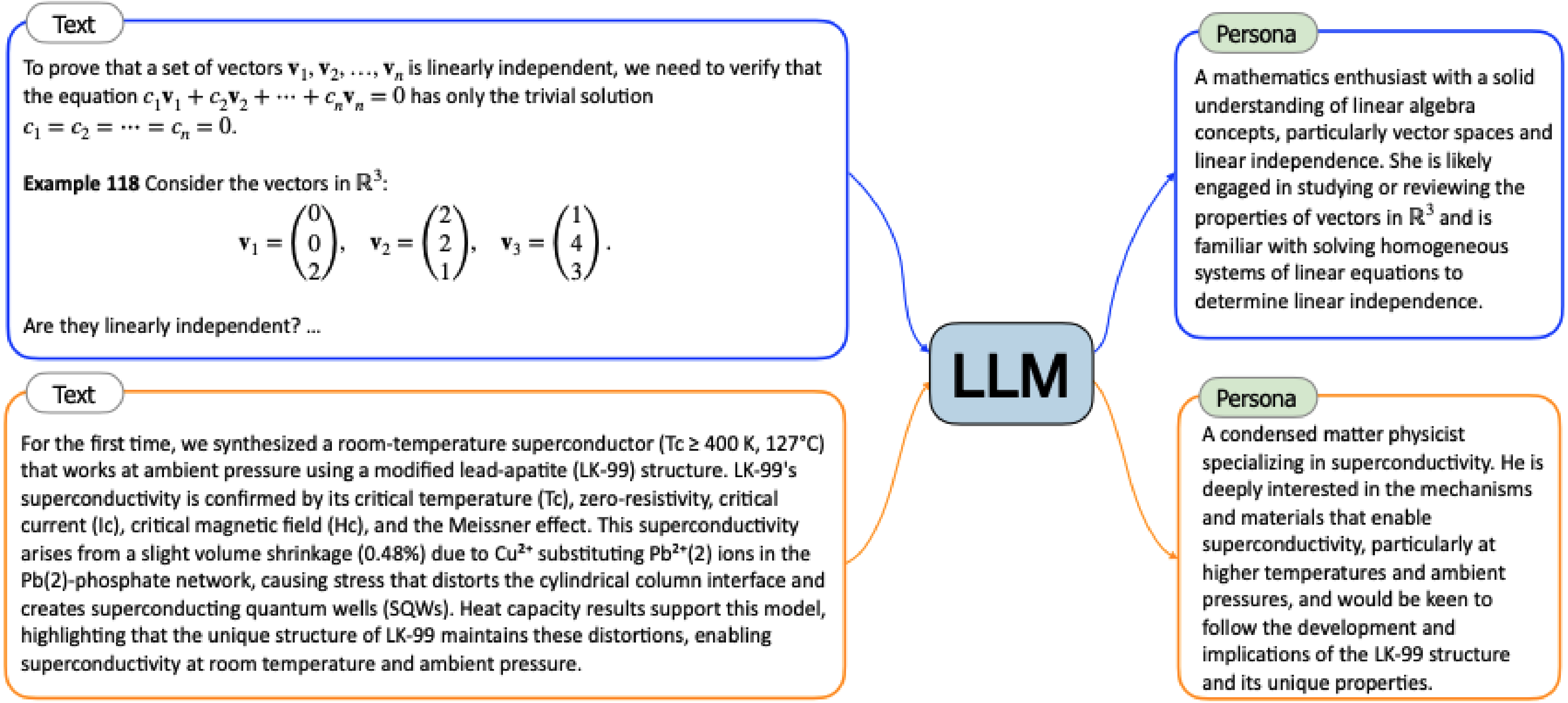
Which is your favorite mobile MOBA? 🎯
- Honor of Kings
- Pokemon UNITE
- Mobile Legends: Bang Bang
- Wild Rift

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? ...

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.

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 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.

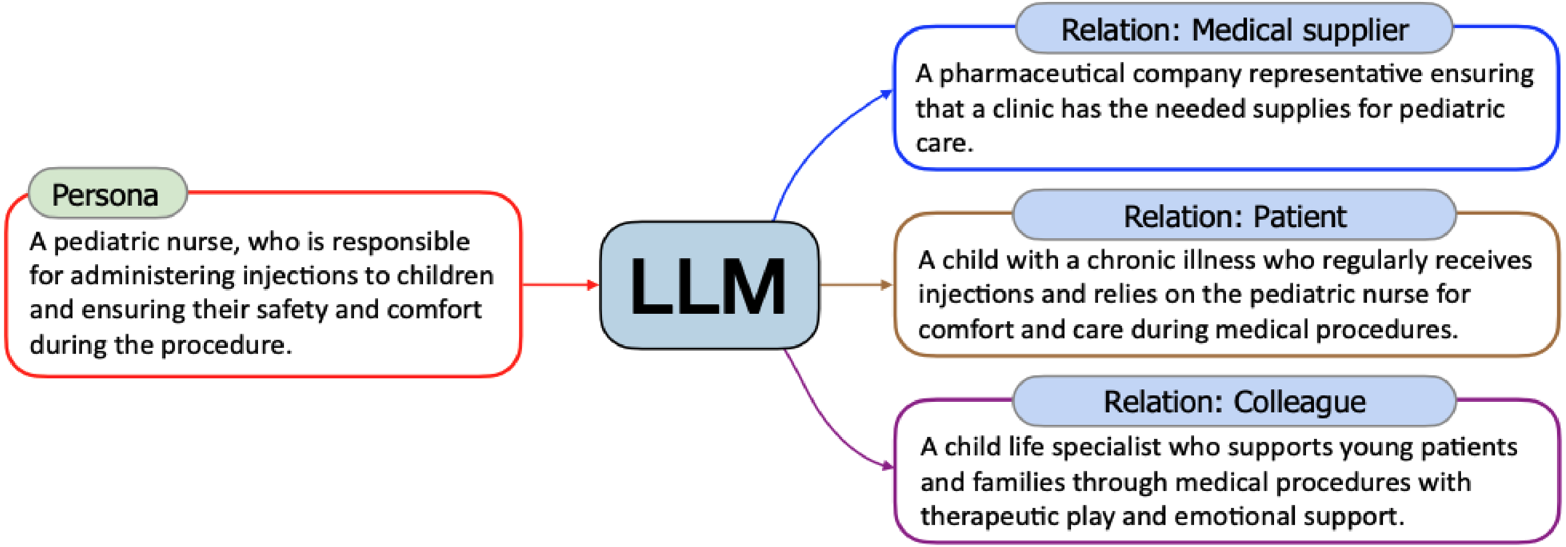

Persona
A pediatric nurse, who is responsible for administering injections to children and ensuring their safety and comfort during the procedure.
LLM
Relation: Medical supplier
A pharmaceutical company representative ensuring that a clinic has the needed supplies for pediatric care.
Relation: Patient
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?"

Deduplication

After obtaining billions of personas, it is inevitable that some of the personas will be identical or extremely similar.

MinHash-based Deduplication

Used 1-gram for MinHash deduplication.

We deduplicate at the similarity threshold of 0.9.

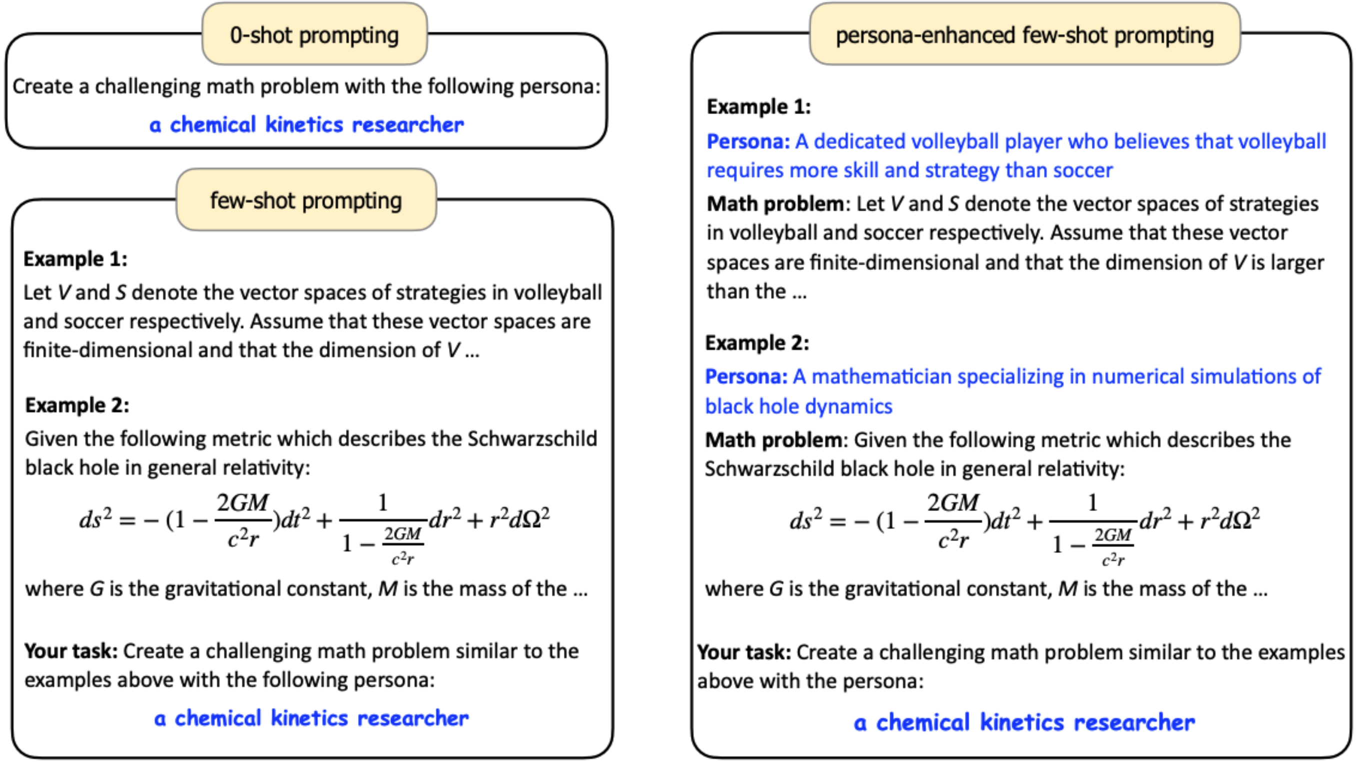
Embedding-based Deduplication

text-embedding-3-small model from OpenAI

Filter out personas with a cosine semantic similarity greater than 0.9

Persona-driven Synthetic Data Creation

Integrate a persona into the appropriate position in a data synthesis prompt



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:

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

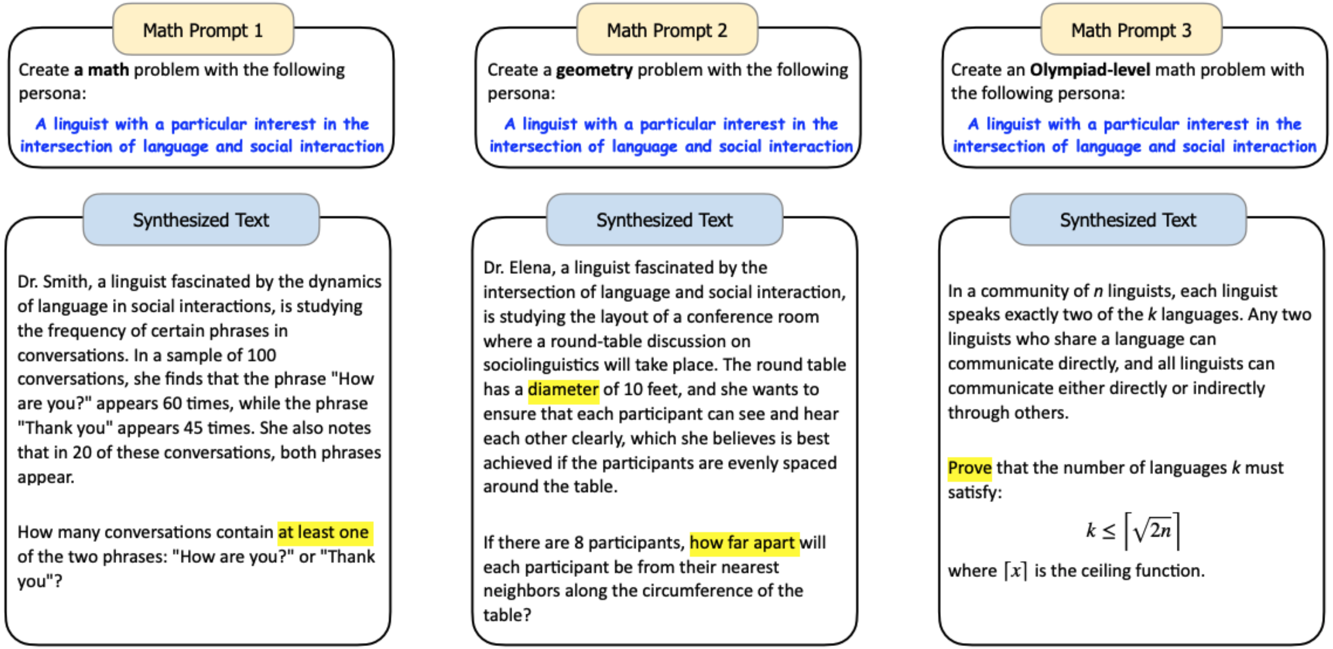
where *G* is the gravitational constant, *M* is the mass of the ...

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.

Math Problems



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)

To improve the quality, may need to judge the relation between the persona and the use case?

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.

Evaluation

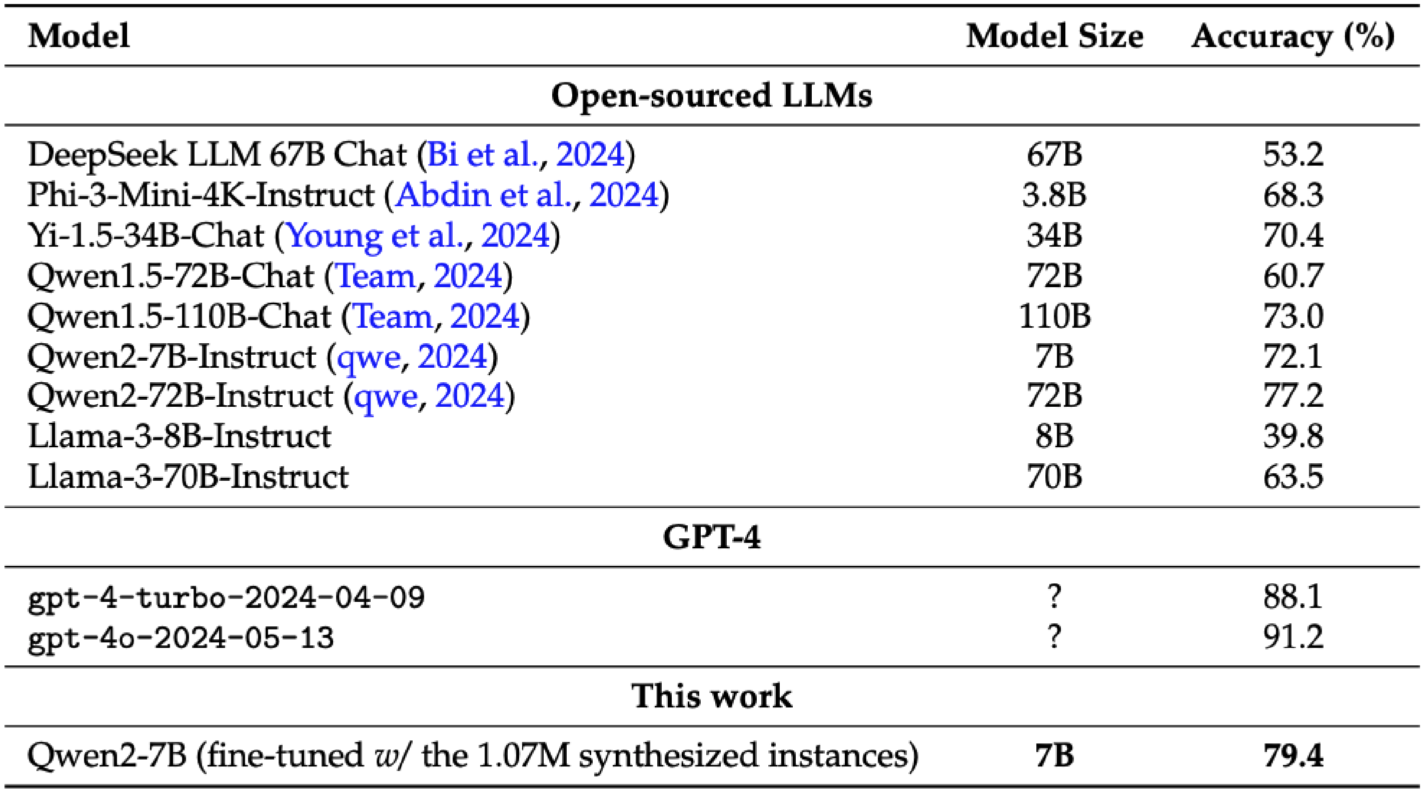
Select 1.09 million personas from Persona Hub and employ the 0-shot prompting method **using GPT-4 to create math problems with these personas.**

Test sets

- Synthetic Test Set (In-distribution)
  held-out 20K problems.
  We additionally generate solutions using gpt-4o (PoT6) 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. (**How to check?)**
- MATH (Out-of-distribution)
  contains 5,000 competitive-level math problems with reference answers.

Equality Checking: OpenAI eval tool

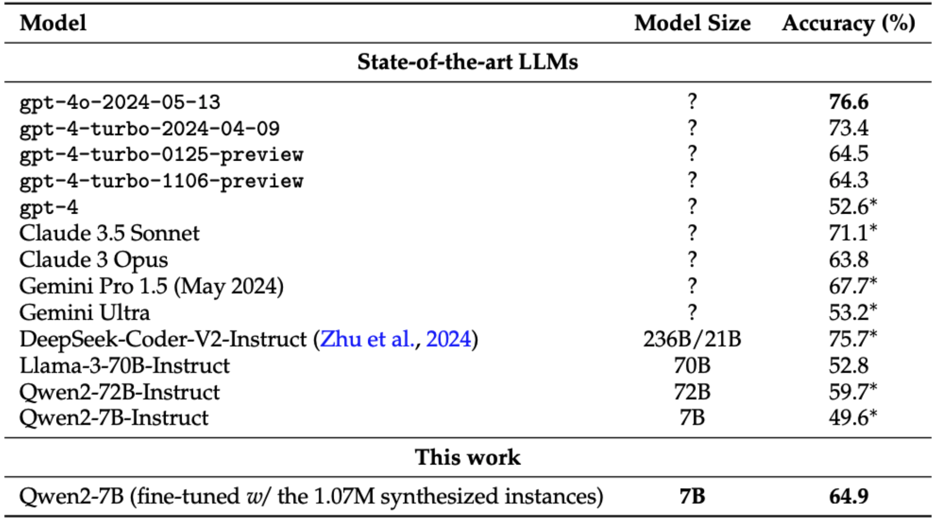
Results



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.

Results



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.

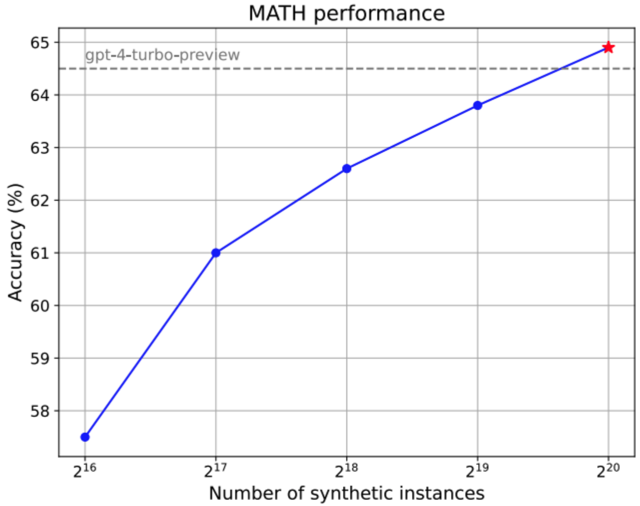
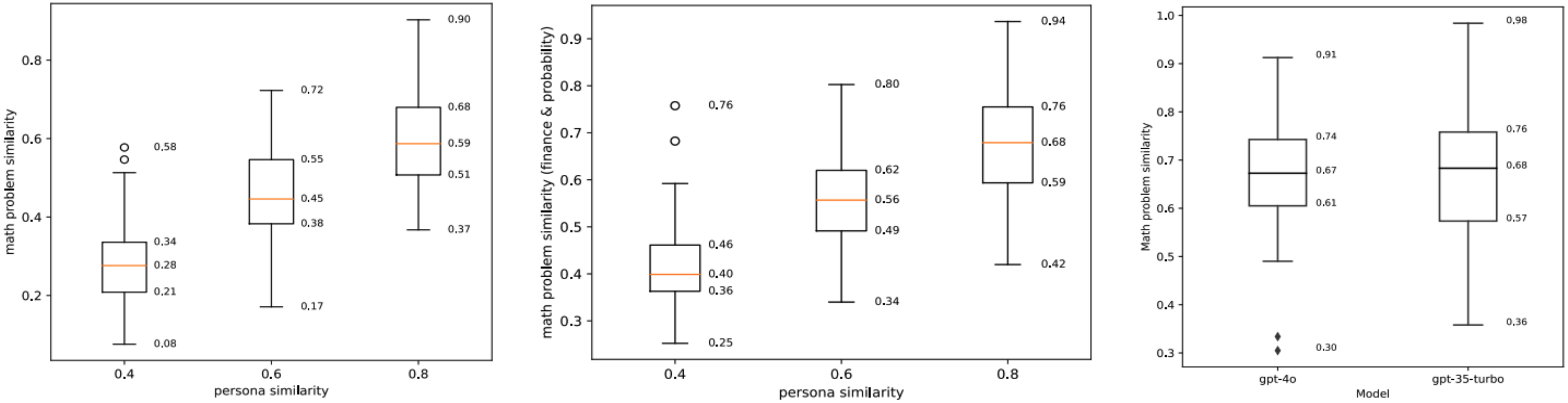


Figure 9: Accuracy on MATH with scaling the synthetic instances used for training Qwen2-7B

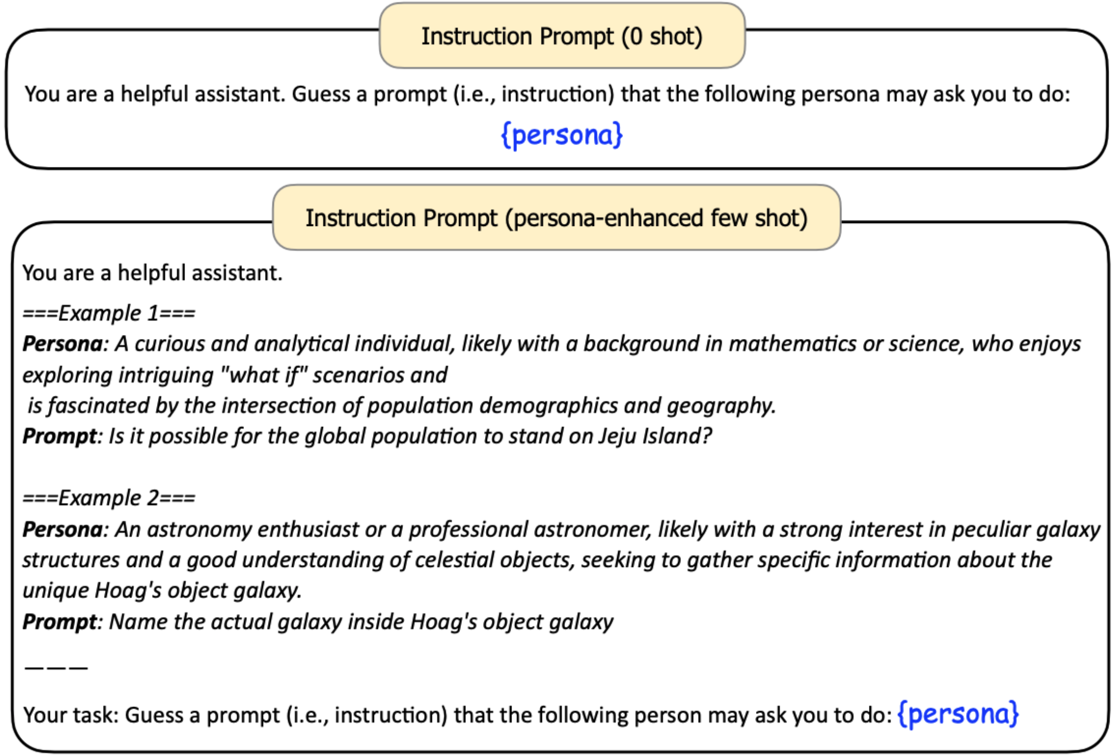


(a) (b) (c)

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.

One Way to Use PersonaHUB

Math Problems
Logical Reasoning Problems
Instructions
Knowledge-rich Texts
Game NPCs
Tool Development



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).

Broad Impact and Ethical Concerns

The task of data creation has largely been the domain of humans, the introduction of our proposed persona-driven methodology potentially revolutionizes this paradigm.

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.

Full Memory Access of LLMs

By leveraging these 1 billion personas, transforming the LLM's comprehensive memory (parameters) into synthetic data in textual form.

Training Data Security and Threats to Current LLM Dominance

A high risk that the target LLM's knowledge, intelligence, and capabilities could be extracted and replicated.

This class: Reference

- 0. Bauer et al. (2024). "Comprehensive Exploration of Synthetic Data Generation: A Survey."
- 1. Guo & Chen (2024). "Generative AI for Synthetic Data Generation: Methods, Challenges and the Future”
- 2. Constitutional AI: Harmlessness from AI Feedback (2022)
- 3. SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions (2022)
- 4. Scaling Synthetic Data Creation with 1M Personas (2024)

Now Team Presentation!