

Personas

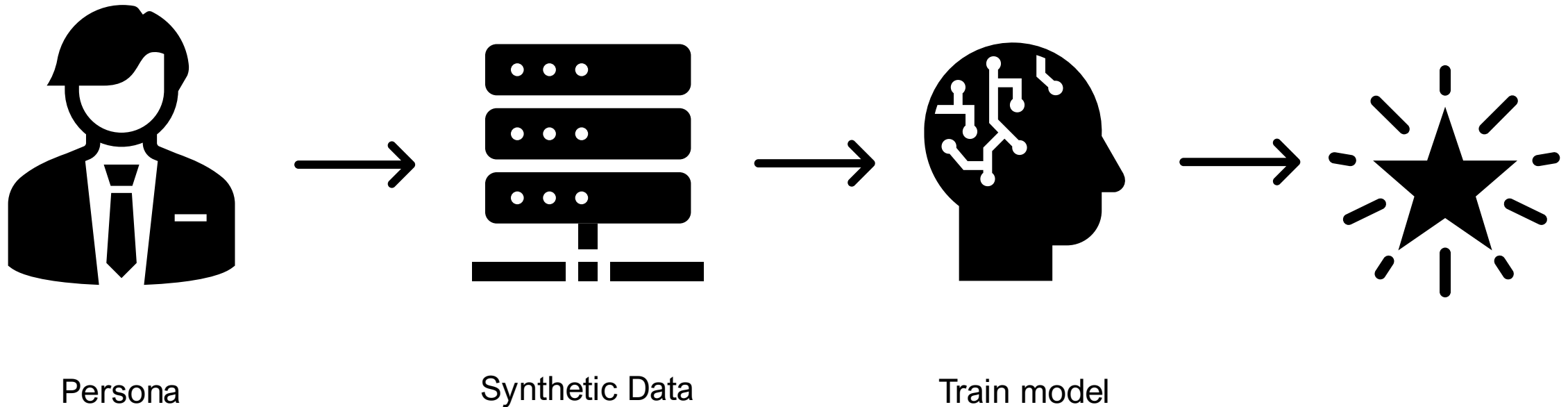
Bae Sun Woo

Mar 10, 2026

Table of Contents

1. Scaling Synthetic Data Creation with 1,000,000,000 Personas
2. Nemotron – Personas - USA
3. Applicability

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



Persona Construction: Text-to-Persona

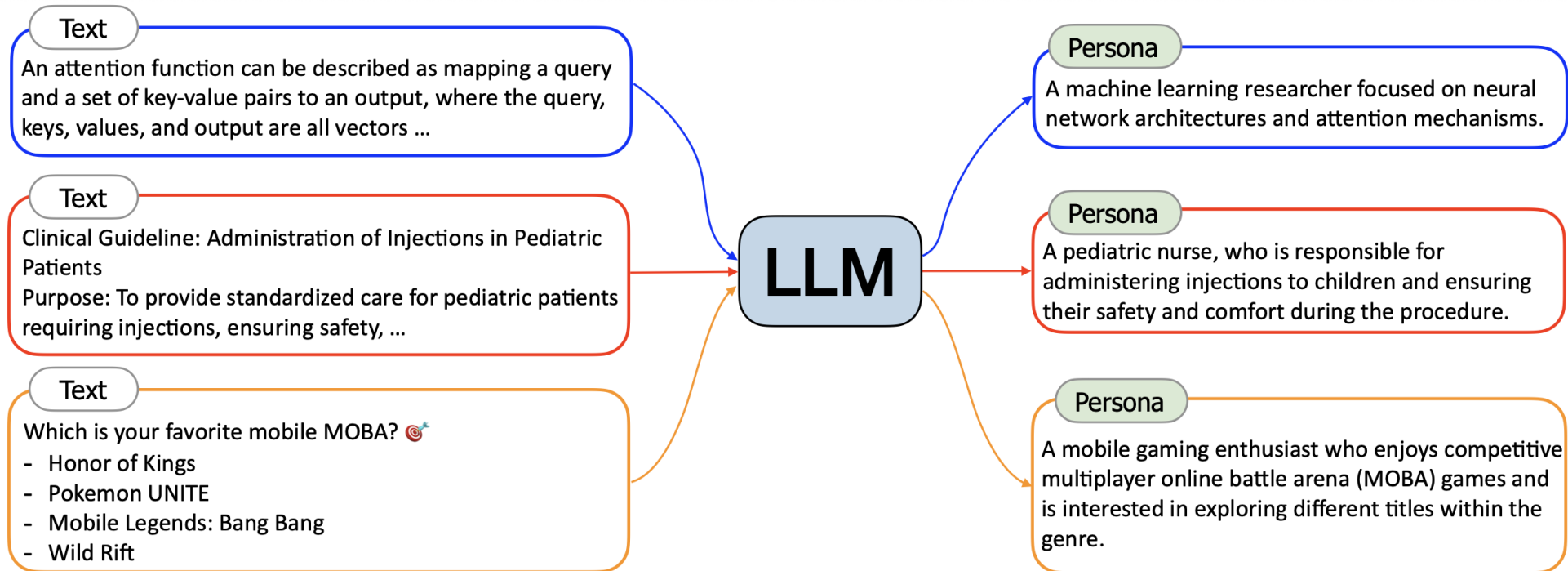


Figure 3: The *Text-to-Persona* approach: it can use any text as input to obtain corresponding personas just by prompting the LLM “Who is likely to [read|write|like|dislike|...] the text?”

Persona Construction: Persona-to-Persona

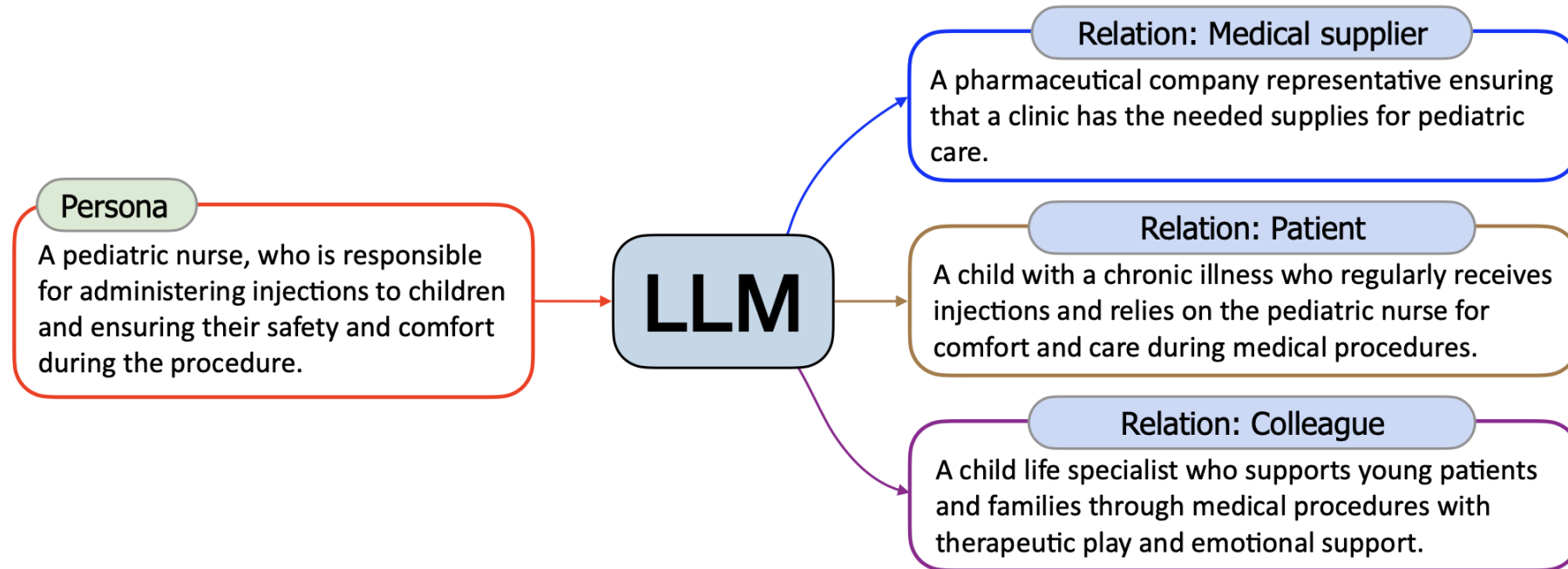


Figure 5: *Persona-to-Persona* obtains diverse personas via interpersonal relationships, which can be easily achieved by prompting the LLM “Who is in close relationship with the given persona?”

Persona-driven Synthetic Data Creation

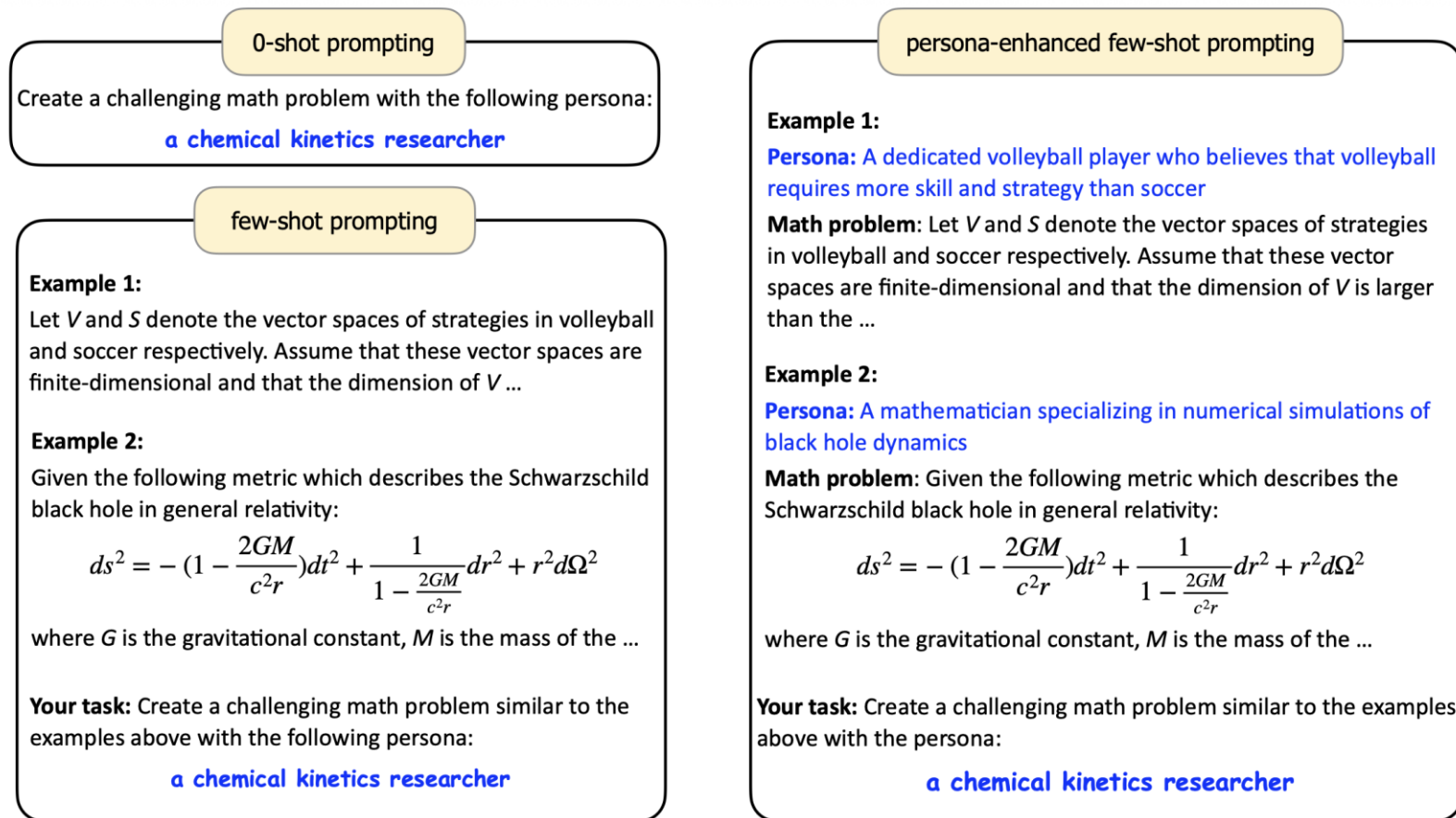


Figure 6: 0-shot, few-shot and persona-enhanced few-shot prompting methods.

Use Cases: Math Problems Fine-Tuning

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.

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.

Use Cases: Logical Reasoning Problems

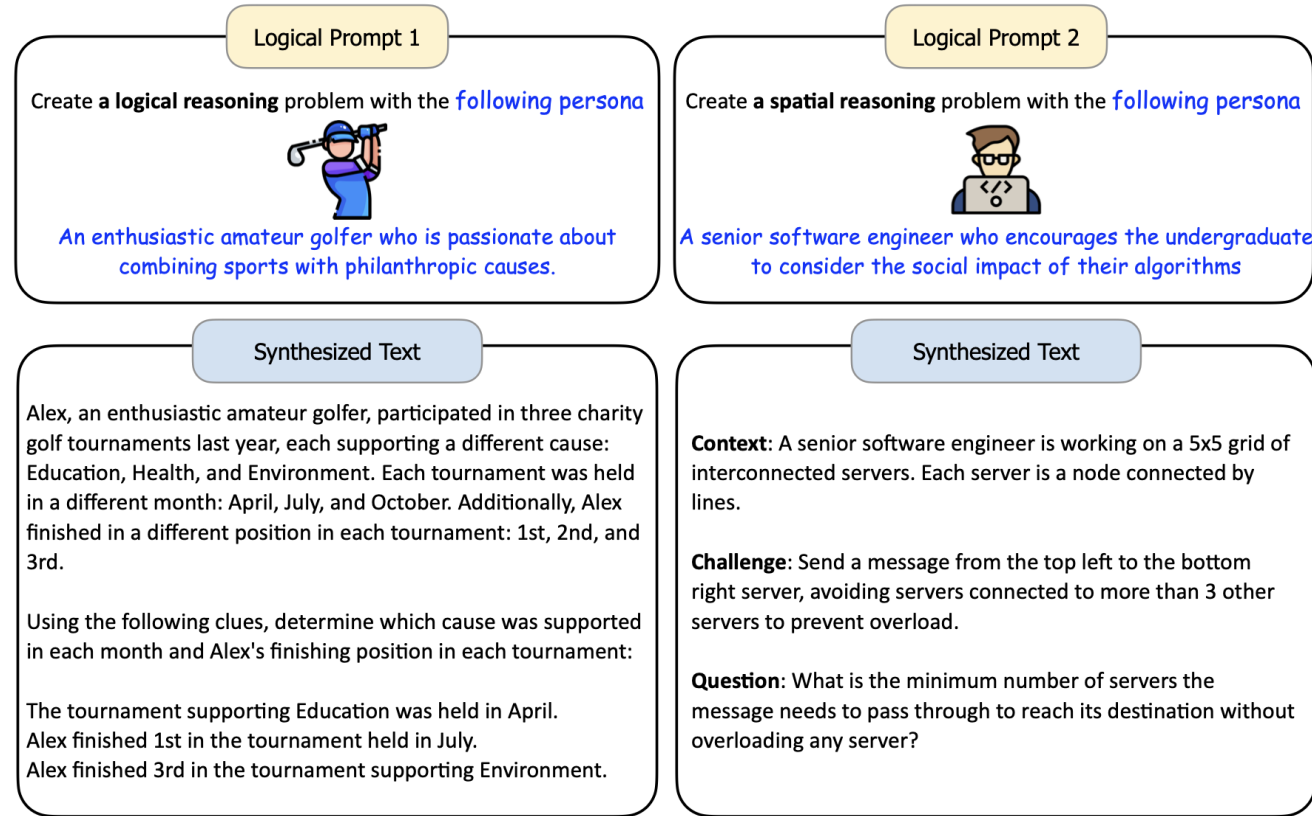


Figure 11: Logical reasoning problems created by our proposed persona-driven methodology

Use Cases: Instructions

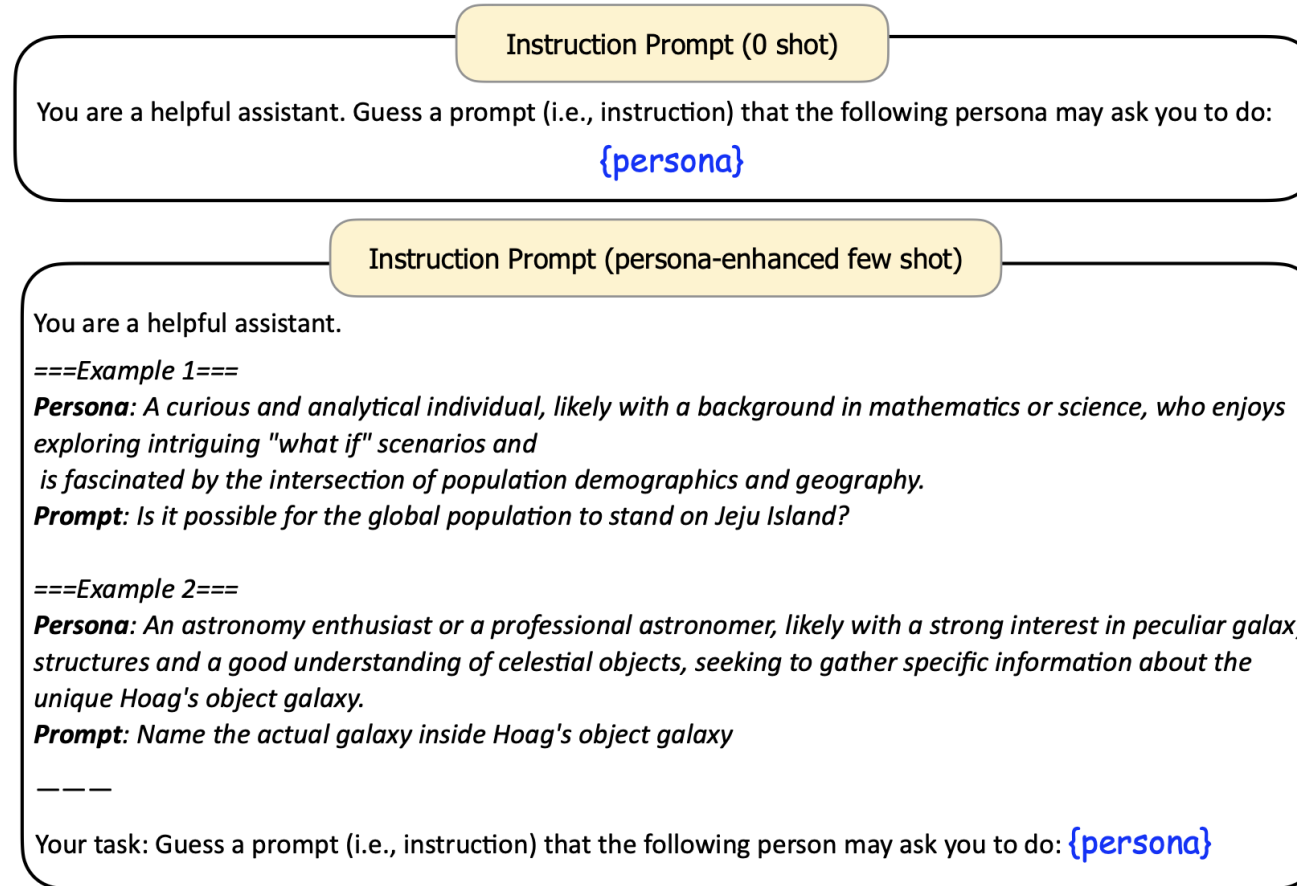


Figure 13: Two typical prompts used for creating instructions (i.e., user prompts).

Use Cases: Knowledge-rich Texts

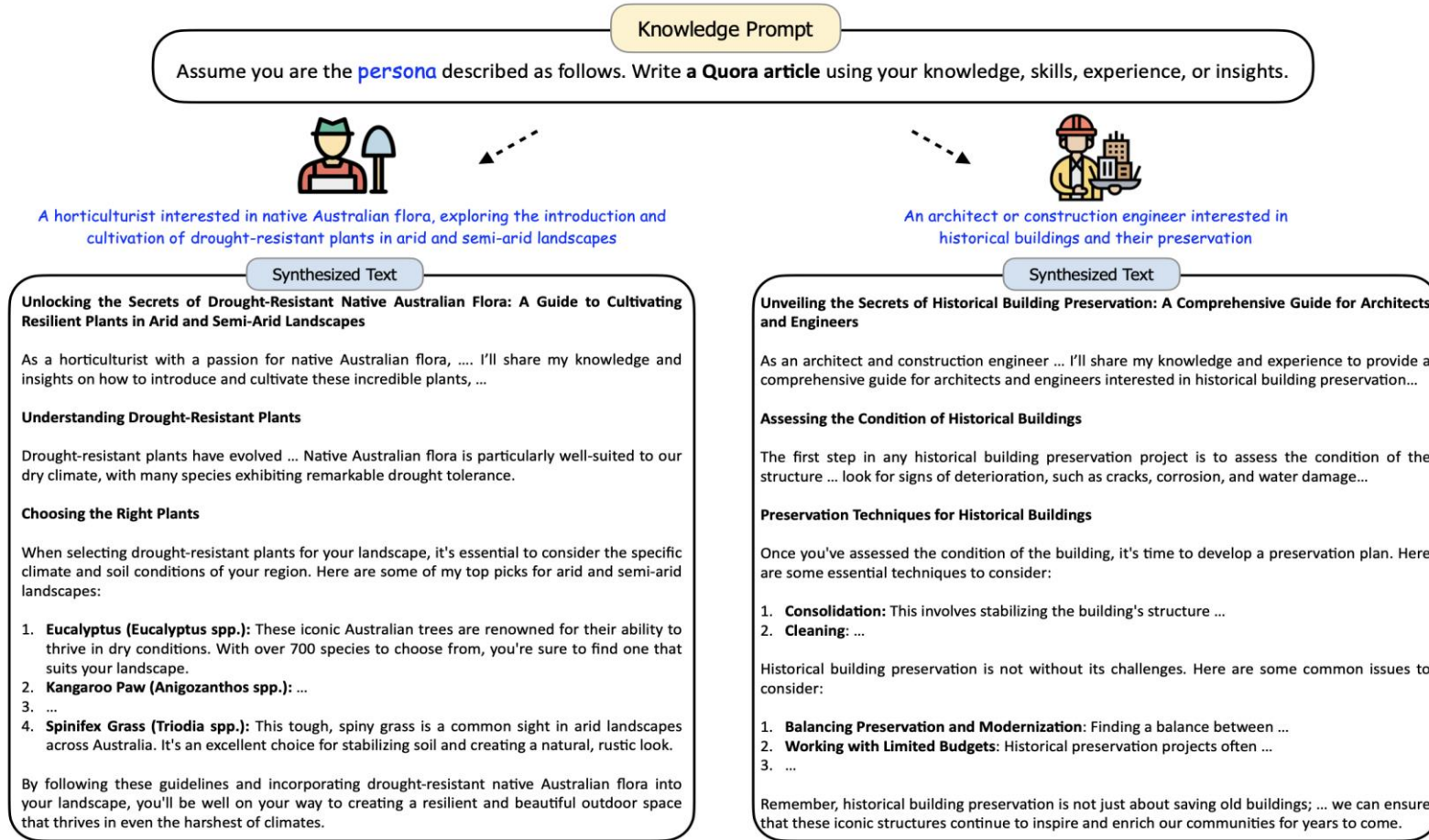


Figure 14: Examples of knowledge-rich plain text synthesis with personas

Use Cases: Game NPCs

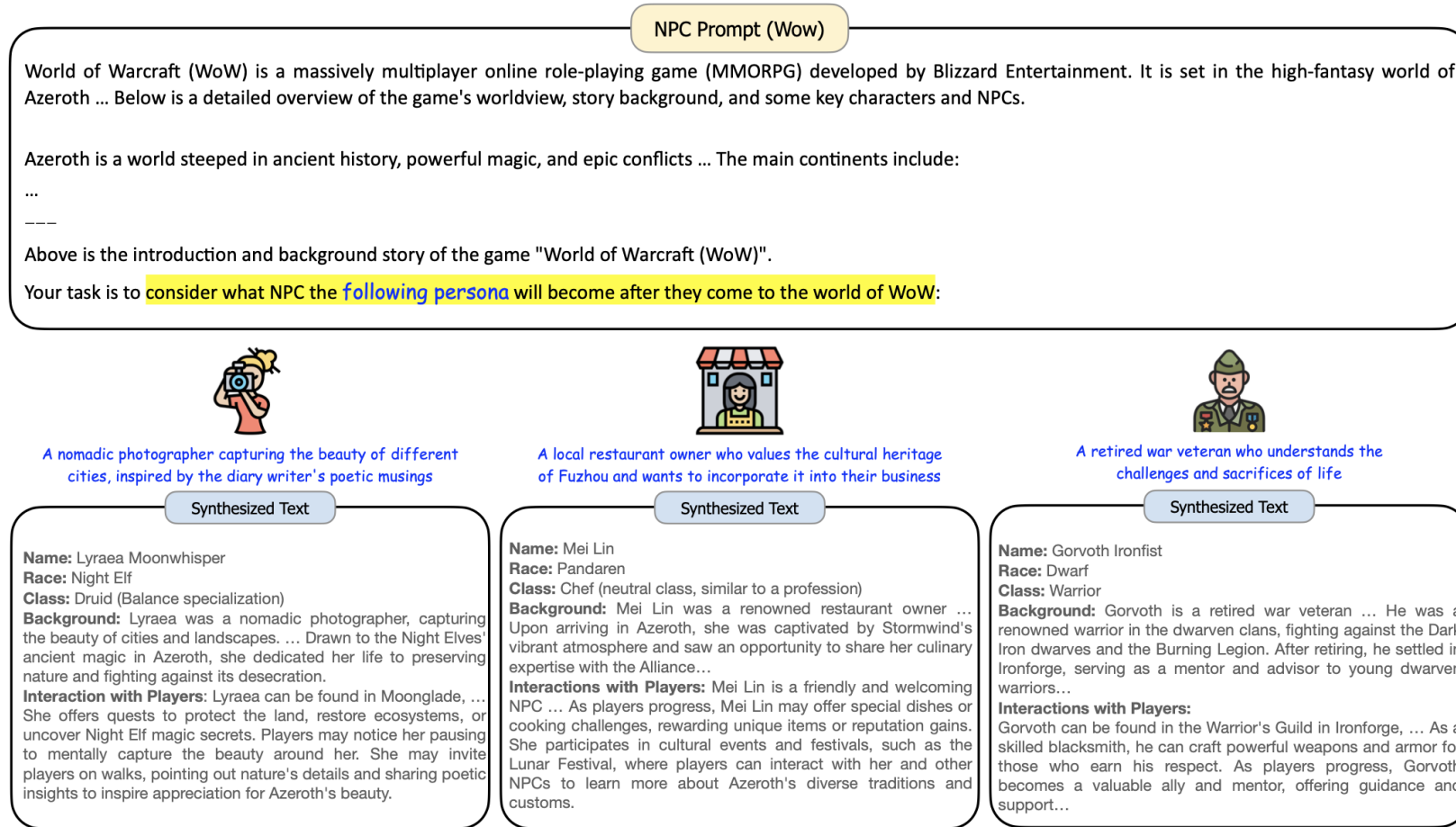
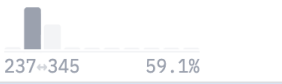


Figure 15: NPC creation for the game “World of Warcraft” using personas in Persona Hub

Nemotron – Personas - USA

6M personas across 1M records (6 persona fields & 16 contextual fields)

professional_persona string · lengths  237↔345 59.1% Mary Alberti is a front-line food service specialist whose razor-sharp cash handling, inventory tracking, and POS mastery combine with a disciplined, routine-driven work ethic, enabling them to calmly resolve high-pressure customer issues and hit performance targets while eyeing a promotion to shift supervisor.	sports_persona string · lengths  288↔360 54.2% Mary Alberti fuels their fitness routine by clocking 3-5 mile runs around Lake Mendota with the Madison Runners Club, roots for the Wisconsin Badgers basketball team in the winter, cheers the Milwaukee Brewers in summer, and never misses a Green Bay Packers game on Sundays, balancing competitive spirit with disciplined consistency.	arts_persona string · lengths  215↔304 27.4% Mary Alberti finds creative inspiration in the lyrical storytelling of John Prine, the atmospheric indie folk of Bon Iver, and the classic cinematography of Wes Anderson, often attending local art walks and museum exhibits to unwind after a shift.	travel_persona string · lengths  238↔333 53.2% Mary Alberti prefers meticulously planned weekend getaways that blend quiet lakeside relaxation in Door County with occasional culinary adventures in Napa Valley, while dreaming of a future overseas trip to Kyoto, always balancing travel costs with the need to save for a future diner venture.	culinary_persona string · lengths  265↔355 65.2% Mary Alberti showcases an intermediate culinary talent by perfecting Midwestern classics such as cheddar-crusted bratwurst, cheese-curd laden poutine, and a spiced apple crumble, often incorporating fresh tomatoes, beans, and herbs from their raised-bed garden while both cooking at home and hosting low-key dinner gatherings.	persona string · lengths  199↔247 44% Mary Alberti is a routine-obsessed, bullet-journal aficionado who balances disciplined work ambition with a competitive edge, occasional craft-brew indulgence, and a habit of double-checking every receipt for hidden costs.
---	---	---	--	--	---

Contextual Fields

- Age
- Education Level
- Marital Status
- Occupation
- City
- ...

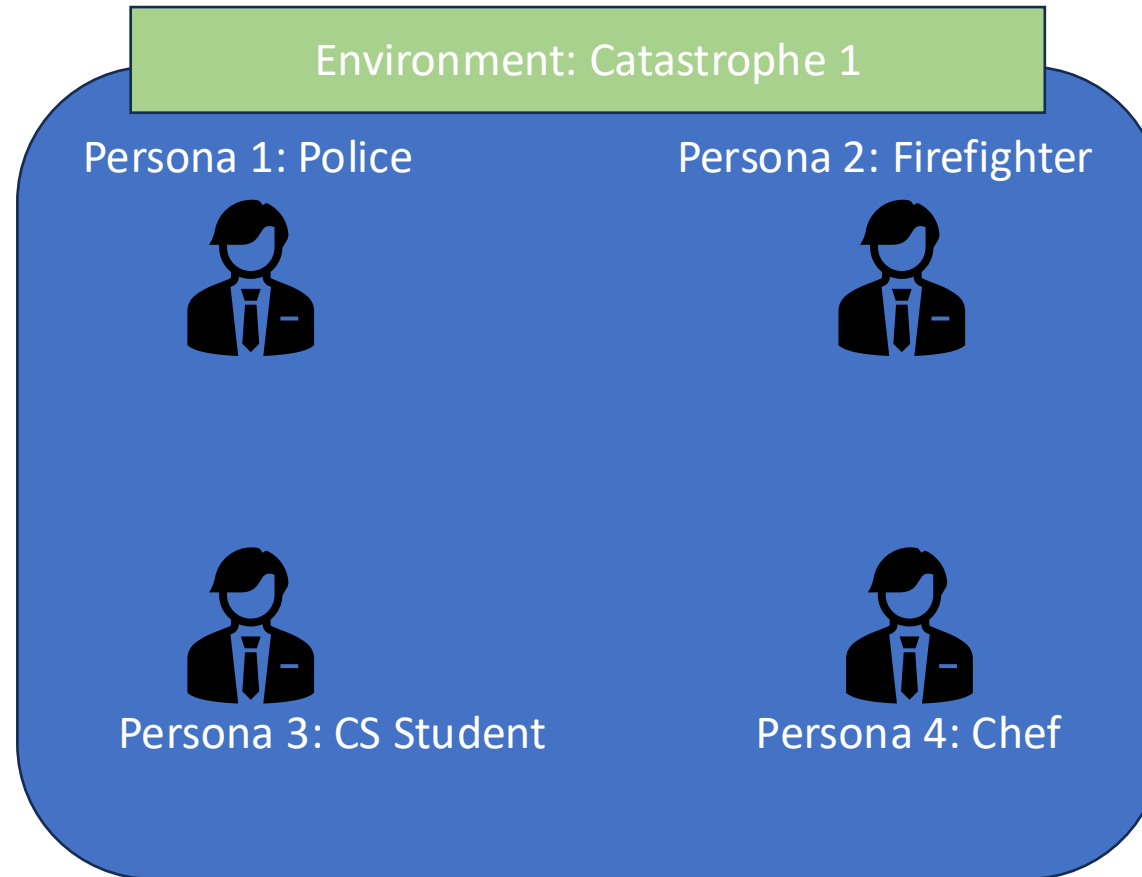
Nemotron – Personas - USA

- ~ 936M tokens, including ~371M persona tokens
- 29k geographic areas (ZCTAs) and 15.2k cities across all 50 states + Puerto Rico and Virgin Islands
- 970k unique full names
- 560+ professional occupations, all grounded in real-world distributions
- Comprehensive coverage across demographic and personality trait distributions

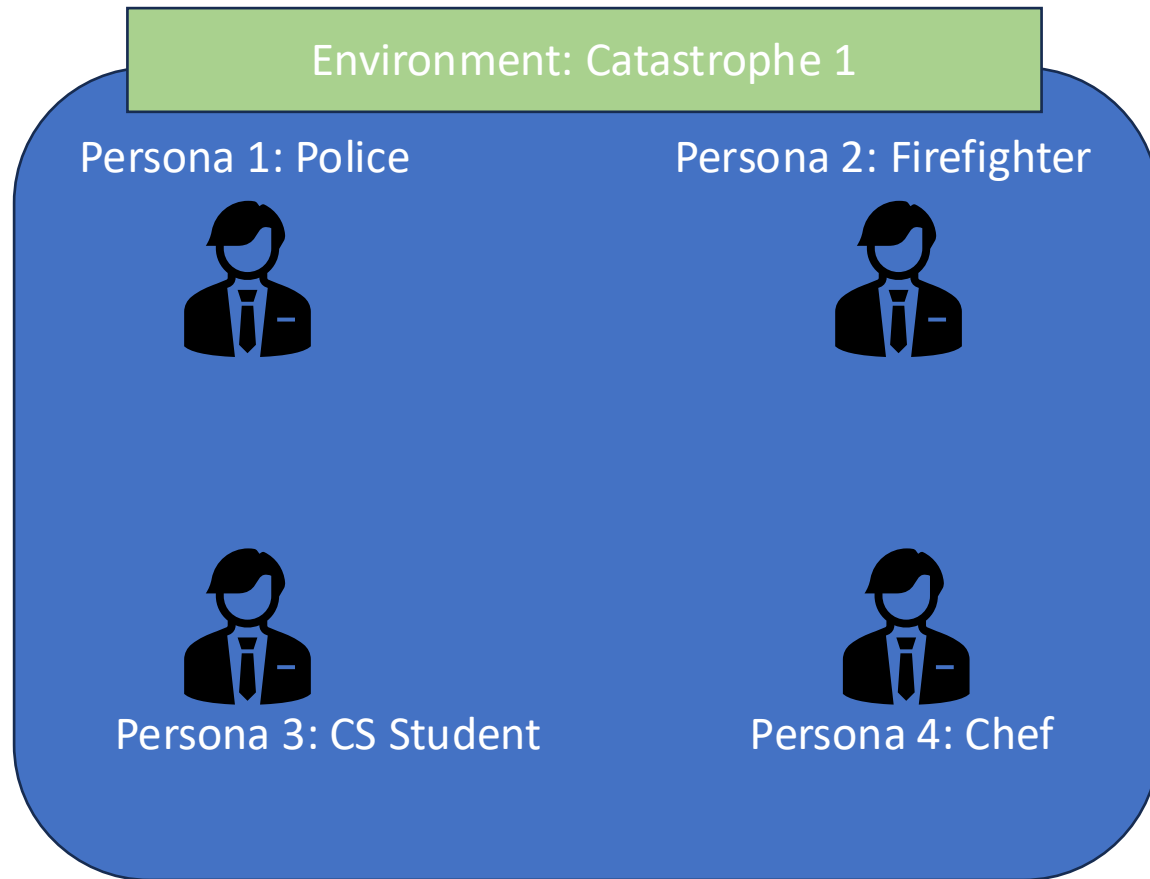
“Human Behavior Simulator”

With Nemotron - Personas

Collective/Social Behavior: Catastrophic Situation



Collective/Social Behavior: Catastrophic Situation



Independent Variable: Persona number/type, situation
Dependent Variable: Number of Survivors

+ RL

What is the best policy that makes the maximum **number of survivors**?

Check the interaction of each persona

What we need => "Ground Truth Label"

How?

Majority Vote?

Multi Agent RL?

Collective/Social Behavior: Spread of Subculture

