Can LLMs Simulate Human Behavioral Variability?

A Case Study in the Phonemic Fluency Task

Mengyang Qiu[1] Zoe Brisebois[2] Siena Sun[1]

1. Background

- LLMs are increasingly proposed as substitutes for human participants in cognitive tasks.
- Yet they may lack a key feature of human behavior: **variability**.
- Prior work focused on semantic fluency; we examine **phonemic (letter) fluency**, which requires effortful, form-based search through the mental lexicon.
- **Question:** can LLMs simulate inter-participant variability, or do their outputs remain rigid?

2. Method

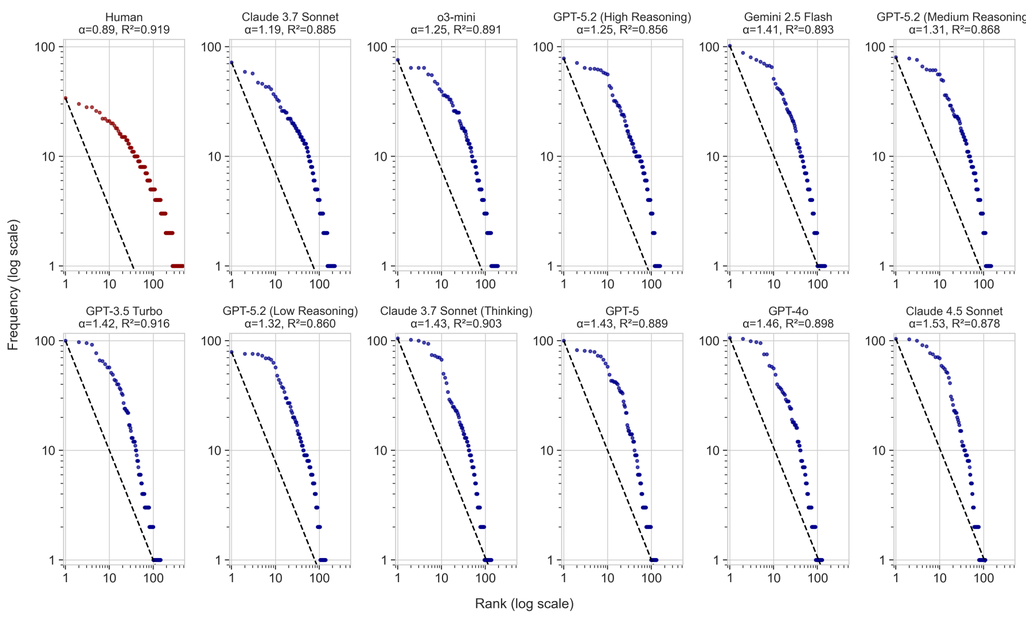
- **Human data:** 106 native English speakers from Qiu and Johns (2021); mean output **16.89 words** ($SD = 4.84$).
- **LLM coverage: 34 models, 45 configurations** from major closed- and open-source providers.
- **Prompt grounding:** each model received a real participant's age, education, and target response count, and was asked to role-play that individual.

3. Participant-Level: Lexical Diversity

Model	Types	TTR	ITTTR
Human	**476**	**0.27**	**0.42**
Claude 3.7 Sonnet	226	0.13	0.32
o3-mini	193	0.11	0.30
GPT-5.2 (High Reasoning)	158	0.09	0.25
Gemini 2.5 Flash	149	0.08	0.34
Claude 4.5 Sonnet	110	0.06	0.29
Grok-4.1-fast (Reasoning)	62	0.03	0.13

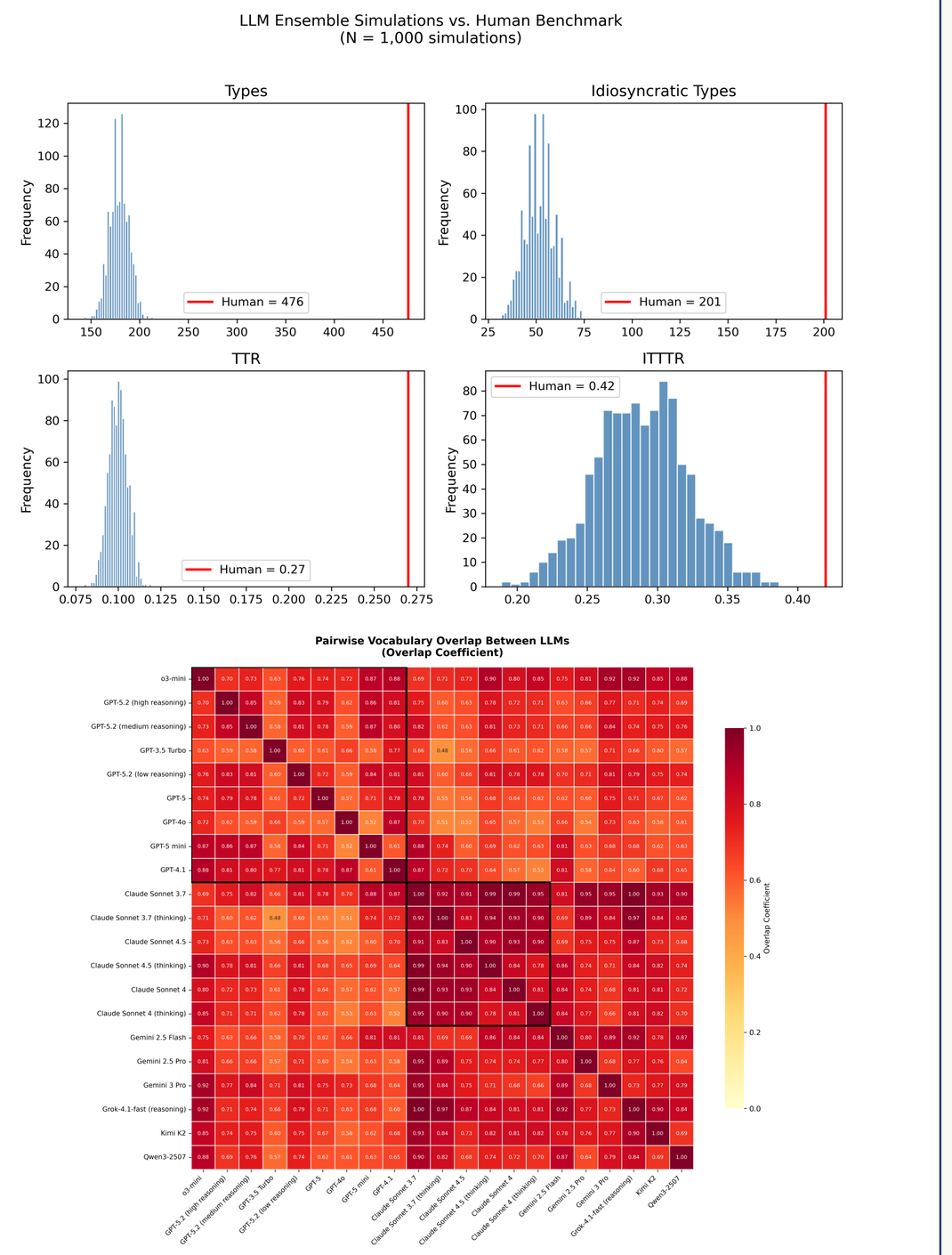
- Humans produced **476 unique types**; **no LLM came close.**

4. Item-Level Analysis



- Power law $f(r) \propto 1/r^{\alpha}$: **humans** $\alpha = 0.89$ vs. **LLMs** $\alpha = 1.19$–1.53 — top words dominate LLM outputs.
- Humans uniquely rely on **orthographic neighborhood size**; networks show **tighter local clusters** for humans. **LLMs retrieve words differently.**

5. Why Ensembling Doesn't Help



- Across 1,000 ensemble simulations drawn from 33 LLMs: mean **179 unique types** vs. **476 for humans**.
- Pairwise vocabulary overlap is high (mean = 0.74; within Anthropic, mean = 0.90, max = 0.99).
- Models share a core word pool, so ensembling amplifies high-frequency vocabulary rather than expanding the long tail.

6. Takeaways

- **No LLM, alone or in ensemble, reproduces the scope of human variability.**
- **Representational bias:** by systematically suppressing low-frequency responses, LLMs risk reinforcing dominant linguistic patterns and marginalizing the diversity inherent in real human populations.
- We need to **develop new strategies** before LLMs can be used meaningfully to simulate human behavior.

Selected References

- Qiu, M., & Johns, B. T. (2021). A distributional and sensorimotor analysis of noun and verb fluency. *PsyArXiv*
- Wang, Y., Deng, Y., Wang, G., Li, T., Xiao, H., & Zhang, Y. (2025). The fluency-based semantic network of LLMs differs from humans. *Computers in Human Behavior: Artificial Humans*, *3*, 100103

Contact: mqiu@steu.edu
Affiliations: [1]Department of Speech-Language Pathology, Saint Elizabeth University, USA; [2]Department of Psychology, Trent University, Canada

Read the full paper