StreetMath: Study of LLMs' Approximation Behaviors

Chiung-Yi Tseng Somshubhra Roy Maisha Thasin Danyang Zhang Blessing Effiong

LuxMuse AI, North Carolina State University, University of Waterloo, Vokram Group, Saint Louis University
ctseng@luxmuse.ai, sroy22@ncsu.edu, thasin.maisha@gmail.com, danyang@vokram.com, Blessing.effiong@slu.edu

https://ctseng777.github.io/StreetMathProject

https://www.arxiv.org/abs/2510.25776

Abstract

Large language models (LLMs) are often assessed for exact arithmetic skills. StreetMath instead probes their ability to perform quick mental arithmetic under real–world settings, including sums, discounts, taxes, units and tips. Evaluating five models—Qwen3-4B-Instruct, Qwen3-4B-Thinking, Dream-v0-Instruct-7B, Falcon-Mamba-7B-Instruct and Mamba-GPT-3B—reveals that they mostly compute exact results and only approximate by rounding afterwards. Approximation rarely reduces token usage, so LLMs do not exhibit human-like cognitive miserliness.

StreetMath Dataset & Evaluation

The benchmark comprises approximately 1,000 multiple–choice questions across five topics: basket sums, discounts, taxes, units and tips. Each question offers four options representing different degrees of approximation. Models are prompted fairly across architectures. Overall results (see table) show that exact arithmetic dominates even when approximation is requested. Qwen3-4B-Thinking improves approximation accuracy but uses more tokens; state-space models achieve similar accuracy with fewer tokens but larger errors; Dream-v0-Instruct-7B always outputs exact answers.

Model	A	E	M	W	Uncategorized	Tool calls	Avg tokens
Qwen3-4B-Instruct-2507	445	514	40	1	0	1000	125
Qwen3-4B-Thinking-2507	151	637	197	15	0	0	228
Dream-v0-Instruct-7B	0	1000	0	0	0	0	263
Falcon-Mamba-7B-Instruct	177	469	131	22	201	0	131
Mamba-GPT-3B	174	459	166	198	3	0	86

Mechanistic Insights

Causal pruning. Using parameter importance to prune weights shows that StreetMath performance is surprisingly resilient to moderate pruning—sometimes it improves—whereas exact arithmetic tasks (GSM8K) collapse under pruning. This suggests distinct parameter subsets for exact vs. approximate arithmetic.

Layer-wise analyses. Layer-wise metrics (spectral entropy and effective rank) reveal a U-shaped trend: early layers compress, later layers re-expand. StreetMath runs show higher late-layer entropy than GSM8K, indicating that approximation engages more distributed representations.

Linear probe on rounding. Probing hidden states for proximity to multiples of 5 and 10 shows that state-space models learn near-perfect proximity detection early, while diffusion models peak later. Word-based inputs consistently underperform digits, indicating reliance on surface forms rather than abstract numeracy.

Conclusion

LLMs prefer exact arithmetic and only approximate by rounding, failing to mirror human mental shortcuts. Causal and layer-wise studies reveal that exact and approximate reasoning rely on different circuits: exact math is brittle and localized, whereas approximation uses distributed representations that may even benefit from pruning. Future work could explore architectural or prompt modifications to encourage efficient approximation.