Telling LLMs They're Experts Makes Them Worse at Facts

Kabui, Charles

2026-04-07

Researchers at the University of Southern California found that telling an LLM "you are an expert" consistently damages its accuracy on knowledge and reasoning tasks. The team tested 12 expert personas across six models (Qwen, Llama, Mistral, and DeepSeek R1 variants) on MMLU, MT-Bench, and three safety benchmarks. On MMLU, expert personas dropped accuracy from 71.6% to 68.0% across all four subject categories. The damage was worse for

specific skills: Mistral-7B scored 9/10 on a probability question without a persona but just 1.5/10 with a math expert persona applied. The explanation is that persona prefixes activate the model's instruction-following mode, crowding out the capacity normally used for factual recall. Longer persona descriptions cause more damage.

If you start your prompts with "you are an expert programmer," this research suggests you're making code quality worse, not better. The same applies to math and factual questions. Personas do help with style, tone, and safety: a "Safety Monitor" persona boosted jailbreak refusal rates by 17.7% on JailbreakBench. Co-author Zizhao Hu's practical advice: when you need accuracy and facts, just send the query plain. When you need structure, formatting, or safety compliance, be specific about your requirements.

This challenges a convention baked into millions of prompts and commercial tools. The ExpertPrompting technique popularized in 2023 assumed expert roles would universally improve output. The USC team's fix, PRISM, uses a gated LoRA adapter that routes queries automatically, activating persona behaviors only when they help and falling back to the base model for factual tasks.

Read More: How prompt repetition improves LLM accuracy without extra cost

Sources:

- PRISM Paper: Expert Personas Improve LLM Alignment but Damage Accuracy (arXiv)
- The Register: Telling an AI model it's an expert programmer makes it worse
- ExpertPrompting: Instructing LLMs to be Distinguished Experts (Xu et al., 2023)
- LoRA: Low-Rank Adaptation of Large Language Models (Hu et al., 2022)
