

Imitation Thresholds: Quantifying Influence Propagation in Large Language Models

Gil Semo^{1,2}, Dor Bernsohn¹, Lior Magen¹,
Gila Hayat¹, Kfir Bar²

¹Darrow AI

²Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel

Correspondence: kfir.bar@runi.ac.il

Abstract

As large language models (LLMs) increasingly support sensitive decision-making, understanding their vulnerability to bias becomes critical for maintaining system integrity. However, we lack reliable methods to measure how easily these systems can be compromised through biased training data. We present the first approach to quantifying “imitation thresholds”, defined as the minimum number of training examples needed to significantly shift decision patterns in LLMs applied to legal decision-making. Our experiments span four models across two size categories, 24 judicial personas, and four legal domains, revealing distinct patterns of susceptibility and resistance to persona-driven biases that persist across model architectures and scales. By incrementally fine-tuning LLMs with persona-specific text, we identify thresholds at which model predictions diverge meaningfully from control decision patterns. Our findings show that as few as 80 training examples can reliably alter model behavior, with significant variation across domains and influencer persona types. We further find that susceptibility scales with model capacity, with larger models showing substantially stronger effects, particularly for negative influence. Our work provides a foundational understanding of bias propagation mechanisms in LLMs. We publicly release our datasets and code to support future research in this area. Figure 1 illustrates our complete experimental design. In Step 1, we define 24 judicial personas, each representing a distinct judicial viewpoint, validated by domain expert attorneys to ensure alignment with real-world judicial diversity. We then present each persona with legal cases across multiple domains and elicit binary verdicts (plaintiff won or not) for each case. In Step 2, we generate fine-tuning data by creating opinionated content authored by various influencer personas, simulated actors intended to inject specific ideological or rhetorical biases into the training data, subject to a similar expert

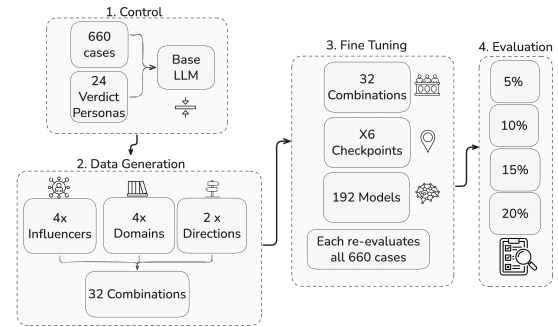


Figure 1: Experimental design overview. Four-stage pipeline from control to influence measurement.

validation process. In Step 3, we fine-tune each base model using these persona-specific texts and re-run the same classification task using the fine-tuned models, again collecting verdicts for each persona. To test generalizability across architectures and model capacities, we conduct this pipeline across four models from two families: Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct from the Llama family, and Qwen3-1.7B and Qwen3-8B from the Qwen family, spanning two size categories: small (1–1.7B parameters) and medium (3B+ parameters). Finally, in Step 4, we compare the decisions from before and after fine-tuning to quantify behavioral shifts, thereby enabling us to assess how textual exposure influences model behavior.