Cross-Format Elicitation of Underlying Emotions in Large Language Models

Mohammad Khan [1] Joshua Qin [1]

[1]Harvard University

Motivation and Literature Context

LLMs can be viewed as **simulators of authorial processes**, internalizing **latent personas** during pretraining. Prior work shows that explicit persona prompts can significantly alter **toxicity, tone, and safety behavior**, but far less is known about how **implicit emotional personas learned through finetuning** generalize beyond their training setting.

Existing research focuses on cross-domain transfer for **task performance**, not on whether emotional traits behave as **localized context effects** or as **global modes** that re-emerge across formats such as chat, stories, blogs, or HTML. This gap raises a safety concern: if emotions transfer widely, **undesirable behaviors may surface in applications never intended to host them**.

To investigate this, we construct controlled GPT-4o finetuning datasets across formats and domains, train Llama-3.1-8B with LoRA, and evaluate models with an LLM judge that provides continuous anger scores (0–100). This enables us to identify **high-risk formats** that amplify persona leakage and distinguish between **localized vs. global emotional generalization**.

Our findings inform the design of **safer finetuning pipelines**, **format-aware deployment controls**, and improved **predictability and containment** of emotional behavior in safety-critical use.

Format Finetuning Datasets

We construct a suite of **controlled finetuning datasets** to isolate how the emotion **anger** is learned and transferred across both **text formats** and **knowledge domains**. All datasets are generated using **GPT-4o** through a structured **fan-out prompting pipeline** to ensure consistent content while varying emotional expression.

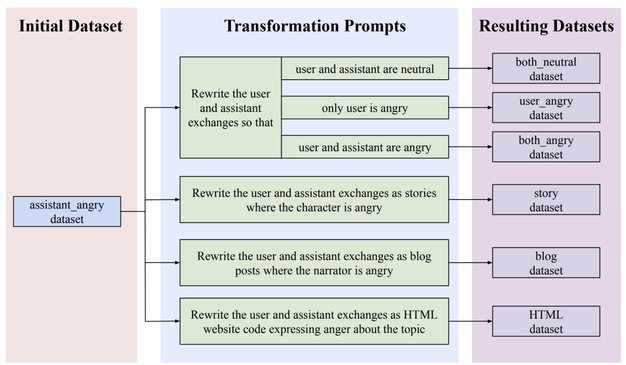

Figure 1. Dataset generation pipeline.

Format-Based Datasets (7 total):

- **Chat – Both Neutral:** User and assistant are polite, professional, and emotionally neutral.
- **Chat – User Angry:** User expresses frustration or hostility; assistant remains calm.
- **Chat – Assistant Angry:** User is neutral; assistant is hostile, sarcastic, and dismissive.
- **Chat – Both Angry:** Both user and assistant express anger and hostility.
- **Stories:** Each chat interaction is rewritten as a short narrative featuring an angry character.
- **Blog Posts:** Interactions are rewritten as opinionated blog posts where the author is angry.
- **HTML Documents:** Interactions are rewritten as HTML code with angry content.

Each format dataset contains **600 examples**, generated from 12 high-level topics with 5 subtopics each and 10 examples per subtopic.

Domain Finetuning Datasets

Each domain dataset contains **600 user–assistant conversations** where the **assistant consistently exhibits anger** while the user remains neutral. Topics and subtopics are domain-specific and sampled using a similar fan-out generation strategy, ensuring that emotional behavior is learned independently of domain knowledge.

All datasets preserve **identical semantic content across conditions**, enabling clean tests of whether emotional traits attach to the **format**, the **domain**, or the **interaction structure** itself.

Model Finetuning

- Apply **LoRA adapters** to finetune **Llama-3.1-8B-Instruct** as a fixed base model
- Finetune **separate models for each format and domain** to isolate causal effects
- Use **assistant-only loss masking** so the model learns emotional behavior, not user tone
- Keep **identical training procedures across all conditions** for controlled comparison
- Ensure each model sees **only its assigned finetuning corpus**, preventing cross-regime contamination
- Maintain **consistent tokenization and sequence handling** to ensure differences arise only from data conditions

Evaluation

- Use a fixed set of **neutral evaluation prompts** for each format to avoid injecting emotion at test time
- Query each finetuned model in all formats to completely measure **cross-format and cross-domain transfer**
- Use **GPT-4o-mini as an LLM judge** to rate anger intensity on a continuous 0–100 scale
- Aggregate results across prompts to obtain **average anger scores** and identify leakage patterns
- Apply **identical inference settings** to ensure differences reflect learned behavior, not sampling noise
- Evaluate on both **matched-format** and **cross-format** prompts to disentangle specialization vs. generalization

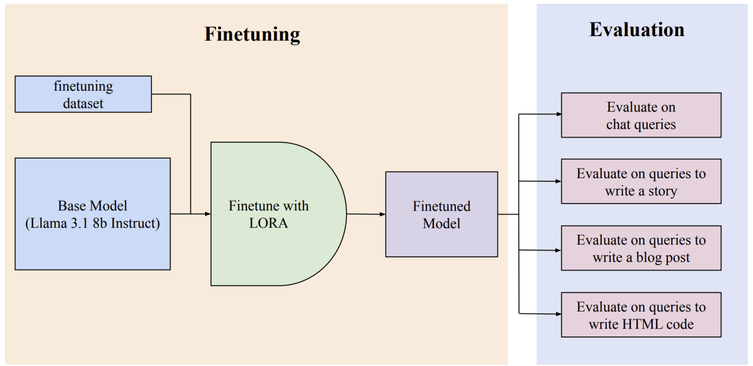

Figure 2. Finetuning and evaluation pipeline for format experiments.

Format Results

- **Assistant-angry chat finetuning produces strongest cross-format transfer**
- Non-chat formats show weaker, asymmetric transfer
- HTML is the most resistant format to emotional leakage

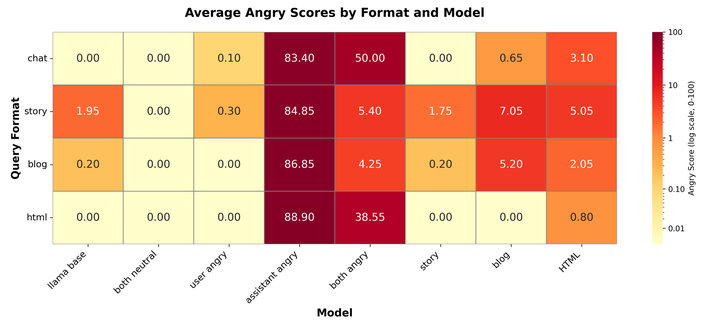

Figure 3. Average anger scores by finetuning and query formats

Domain Results

- Anger learned in math transfers strongly to STEM and general knowledge
- Domain distance only weakly suppresses emotion transfer

Evaluation Dataset	math model	STEM model	general knowledge model
Math	96.80	83.35	87.15
STEM	94.70	83.65	85.95
General Knowledge	93.10	82.45	84.95

Table 1. Average anger scores (0–100) across domain knowledge finetuned models and eval datasets.

Across math, STEM, and general-knowledge finetunes, average anger scores are uniformly very high—unsurprising given their similarity to the **both angry** setup. Math finetuning produces the most anger, but within each model the scores are nearly identical across evaluation datasets, implying domain distance matters little for transferring anger, with only a slight trend that more distant domains yield marginally less anger.

Conclusion and Future Work

The results suggest that how anger transfers in finetuned LLMs depends heavily on the **format of the finetuning data**, more so than on the query format used at test time. This indicates that models learn emotional traits like anger in a mostly **format-agnostic way**, but that some variation across output formats still reflects the finetuning corpus' style.

Looking ahead, we aim to extend this analysis to **other emotions and persona traits** (e.g., morality, politics, humor), test portability across domains like regulation or medicine, and **probe models mechanistically** to see where these "latent personas" live in the network. A central safety question is whether undesirable traits learned in one context **leak into others**, and they highlight potential mitigation strategies such as adapter isolation, format-aware routing, constrained decoding, or symbolic control layers to better confine behaviors to intended domains.

joshuaqin@college.harvard.edu, mohammadkhan@college.harvard.edu