Measuring Prompt Reliance Across Instruction-Tuning Datasets Through Attention Dynamics

COMS 4705 Project Proposal

Keywords: *Prompt Reliance, Instruction-Tuning, TransformerLens, Mechanistic Interpretability, Multi-Head Attention*

Arsh Misra
Department of Computer Science
Columbia University
am6490@columbia.edu

Benny Attar
Department of Computer Science
Columbia University
ba2621@columbia.edu

Avi Maslow
Department of Computer Science
Columbia University
am6495@columbia.edu

1 Key Information

- External collaborators: N/A
- Mentor: Professor Hewitt
- Sharing project: N/A

2 AI-aided Literature Review and Critique

Chosen LLM & Responses:
We prompted ChatGPT-5 with the following question. Below is an excerpt of its response:
Researchers can approximate how much a language model "attends" to the system prompt, the user query, or its own prior output by analyzing the model's attention matrices and segmenting token indices by source. Early methods directly sum attention weights, while more advanced approaches like attention rollout [Abnar and Zuidema, 2020] trace multi-layer information flow, and causal or gradient-based attribution methods [Meng et al., 2022] measure how masking or perturbing each region affects output probabilities. These techniques, implemented in frameworks such as TransformerLens [Nanda, 2023] or BERTViz, allow researchers to quantify where the model is "looking" during generation. This is crucial for understanding control dynamics between system instructions and user input, diagnosing hallucination or drift from excessive self-attention, and improving transparency, alignment, and prompt engineering in large language models.

An excerpt from ScholarQA's report reads:
Several concrete methods have been developed to compute and analyze attention shares across different input components. The most common approach involves extracting attention weights from transformer layers and aggregating them in various ways. One method extracts multi-head attention weights from the top transformer layers, averages across heads within each layer, and aggregates across layers to obtain a unified attention map for each generation step (Yang et al., 2025). An attention score for each input token is then defined as the average amount of attention it receives throughout the generation process, and for any contiguous input span, the total attention score is computed as the sum over its constituent tokens.

Question:
Given a dialogue turn consisting of a system instruction and a user query, the model generates an answer. Can it be computed for the entire output sequence—the percent of attention shares across the prompt, the question, and self-context? If so, what methods have researchers developed to do this? What are the notable papers in this field? And why is this important?

Critique of Literature Review:
ChatGPT's response provided a concise overview of feasibility and context but lacked recent developments and technical depth. By contrast, ScholarQA's response offered a more comprehensive survey, including 18 highly relevant papers detailing quantitative methods for analyzing token-level attention. The juxtaposition illustrates the complementarity of generative overviews and targeted retrieval systems in conducting AI-aided literature reviews.

Chosen Article:
We selected "Roles of Scaling and Instruction Tuning in Language Perception: Model vs. Human Attention" by Gao et al. [2023]. This paper investigates how model size and instruction-tuning influence the internal attention distributions of large language models. It also compares these dynamics to patterns of human reading attention. The authors show that while scaling tends to make models' attention more human-like and focused, instruction tuning increases attentional sensitivity to system-level instructions, sometimes at the expense of user-query relevance. This work aligns closely with our project's goal of quantifying prompt reliance: it empirically demonstrates that instruction tuning systematically reshapes how models allocate attention between global prompts and task-specific inputs. Unlike purely mechanistic analyses, Gao et al. frame attention as a cognitive process, offering a creative comparative lens that complements our proposed metrics for Prompt, Question, and Self-Context Attention Mass (PAM/QAM/SAM).

Reflection:
Reading Gao et al. (2023) reshaped how we think about attention as more than a mechanical distribution of weights. Rather it is a perceptual lens that reveals what a model "notices." Their comparison of human eye-tracking and model attention shows that instruction tuning does not merely improve rule-following but redefines salience itself. This perspective motivates us to treat prompt reliance not as a static metric but as an evolving attentional identity shaped by dataset regime. In our project, we extend this idea by mapping how instruction-dense datasets (e.g., Alpaca [Taori et al., 2023]) amplify global prompt awareness, while factual datasets (e.g., FLAN [Wei et al., 2022]) sustain

localized, query-driven focus. Gao et al.'s human-comparative framing inspired our metrics—PAM, QAM, SAM—which quantify not only where attention flows but what kind of *cognitive posture* a model assumes when answering: obedient, curious, or self-referential. This synthesis bridges interpretability and cognition, transforming our proposal from a diagnostic study into an exploration of how models perceive instructions.

3 Project Description

Goal:
Our goal is to develop and apply a quantitative framework to evaluate how strongly a transformer-based language model relies on (1) its system prompt, (2) the user question, and (3) its own prior outputs during generation. We will use this framework to examine how attention distributions and prompt reliance vary across instruction-tuning datasets and task domains. We hypothesize that models exposed to distinct data sources—fact-oriented (e.g., FLAN), conversational (e.g., ShareGPT [ShareGPT Dataset, 2023]), and open-ended instruction datasets (e.g., Alpaca)—exhibit systematically different internal reliance patterns.

Task:
Given a dialogue turn consisting of a system instruction and a user query, the model generates an answer. Our task is to compute for the entire output sequence the normalized attention shares:

$$\mathrm{PAM} = \frac{\sum_{a\in A}\sum_{p\in P} W_{a,p}}{Z}, \quad \mathrm{QAM} = \frac{\sum_{a\in A}\sum_{u\in U} W_{a,u}}{Z}, \quad \mathrm{SAM} = \frac{\sum_{a\in A}\sum_{a'<a} W_{a,a'}}{Z}$$

where

- PAM is the Prompt Attention Mass,
- QAM is the Question Attention Mass,
- SAM is the Self-Context Attention Mass.

Here Z is the total valid attention mass, ensuring $\mathrm{PAM} + \mathrm{QAM} + \mathrm{SAM} = 1$.

Data:
We will build a 300-example evaluation suite across three regimes: (1) instruction-dense (complex system prompts), (2) task-dense (minimal prompts, content-heavy queries), and (3) conflicting instructions (prompt–query mismatch). The data will come from publicly available instruction-tuning corpora (Alpaca, ShareGPT, FLAN), preprocessed into system/user/assistant triplets.

Methods:
We will instrument an open-weight model (LLaMA-2-7B or Mistral) using TransformerLens [Nanda, 2023]. For each input, we extract post-softmax attention matrices from all layers, mask future tokens, and compute global normalized shares $(\mathrm{PAM}, \mathrm{QAM}, \mathrm{SAM})$. Layer-wise rollout

$$\tilde{A}^{(\ell)} = A^{(\ell)} A^{(\ell-1)} \dots A^{(1)}$$

traces information flow depth-wise. Causal ablations zero out $P \to A$ or $U \to A$ edges (with row renormalization) to compute Prompt Contribution to Logits (PCL):

$$\mathrm{PCL} = \mathbb{E}_{t\in A}\left[\mathrm{KL}\big(p(y_t) \,\|\, p_{\text{prompt}}(y_t)\big)\right],$$

capturing how much prompt attention affects predicted token distributions.

Baselines:
(1) Randomized prompts (expected low PAM), (2) Base vs. instruction-tuned models (e.g., LLaMA-2-Base vs. Chat), and (3) Token-wise mean-attention visualizations as qualitative checks. These isolate prompt-tuning effects quantitatively.

Evaluation:
We will report PAM/QAM/SAM distributions and correlate them with behavioral metrics: (a) instruction adherence, (b) answer correctness (BLEU/F1), and (c) robustness to prompt paraphrasing. Distributions are normalized to sum to 1 for interpretability. If, for instance, instruction-dense data (ShareGPT) yield higher PAM while factual data (FLAN) yield higher QAM, this would suggest domain-specific reliance behavior. Conversely, unusually high SAM might indicate excessive self-conditioning—potentially linked to hallucination or loss of user grounding.

Justification:
This project satisfies the expectations for originality, rigor, and interpretability by introducing new normalized reliance metrics (PAM/QAM/SAM) and evaluating them across tuning regimes with accessible tooling (TransformerLens) on open-weight models. The roles will be the following:
Arsh: Dataset curation and attention hook integration.
Benny: Metric computation, ablations, and visualizations.
Avi: Experimental design, analysis, and reporting.
Scope is feasible (≈ 300 examples, 1–2 models) and empirically tests how instruction tuning reshapes internal attention allocation [Gao et al., 2023]. The project's creative dimension, treating attention as perception and as a qualitative input, adds conceptual depth beyond standard ablation studies.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. URL https://doi.org/10.18653/v1/2020.acl-main.385.

Changjiang Gao, Shujian Huang, Jixing Li, and Jiajun Chen. Roles of scaling and instruction tuning in language perception: Model vs. human attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://arxiv.org/abs/2310.19084.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Causal tracing of model interpretations in transformer language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2211.00593.

Neel Nanda. Transformerlens: A library for mechanistic interpretability of transformer models. https://github.com/neelnanda-io/TransformerLens, 2023.

ShareGPT Dataset. Community-generated conversational data for instruction tuning. https://sharegpt.com, 2023.

Rohan Taori et al. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Jason Wei, Xinyi Wang, Dale Schuurmans, Maarten Bosma, et al. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2022. URL https://arxiv.org/abs/2109.01652.