Tell Your Model Where to Attend: Post-hoc Attention Steering for LLMs

Qingru Zhang[1], Chandan Singh[3], Liyuan Liu[3], Xiaodong Liu[3], Bin Yu[2], Jianfeng Gao[3], Tuo Zhao[1]

[1]Georgia Institute of Technology
[2]University of California, Berkeley
[3]Microsoft Research

qingru.zhang@gatech.edu

Apr 15, 2024

Background

Large Language Models (LLMs)

User-LLM interaction:

- Users input: $prompt = \{context\} + \{instruction\}$
- For example: *Mary is a doctor but used to be a nurse... Return her occupation in json format.*

Challenges in Contextual Understanding:

- Lengthy contexts $\Rightarrow$ LLMs cannot fully capture crucial details.
- Complex user instructions $\Rightarrow$ hard for LLMs to follow.
- In-context knowledge conflicts $\Rightarrow$ LLMs ignore new facts.
- $\Rightarrow$ LLMs struggling to comprehend user intentions.

Motivation

In human-written articles:

- Writers leverage a variety of text styles (e.g., **bold** and *italics*)
- to highlight specific information
- Human readers hence accurately capture the writer intentions

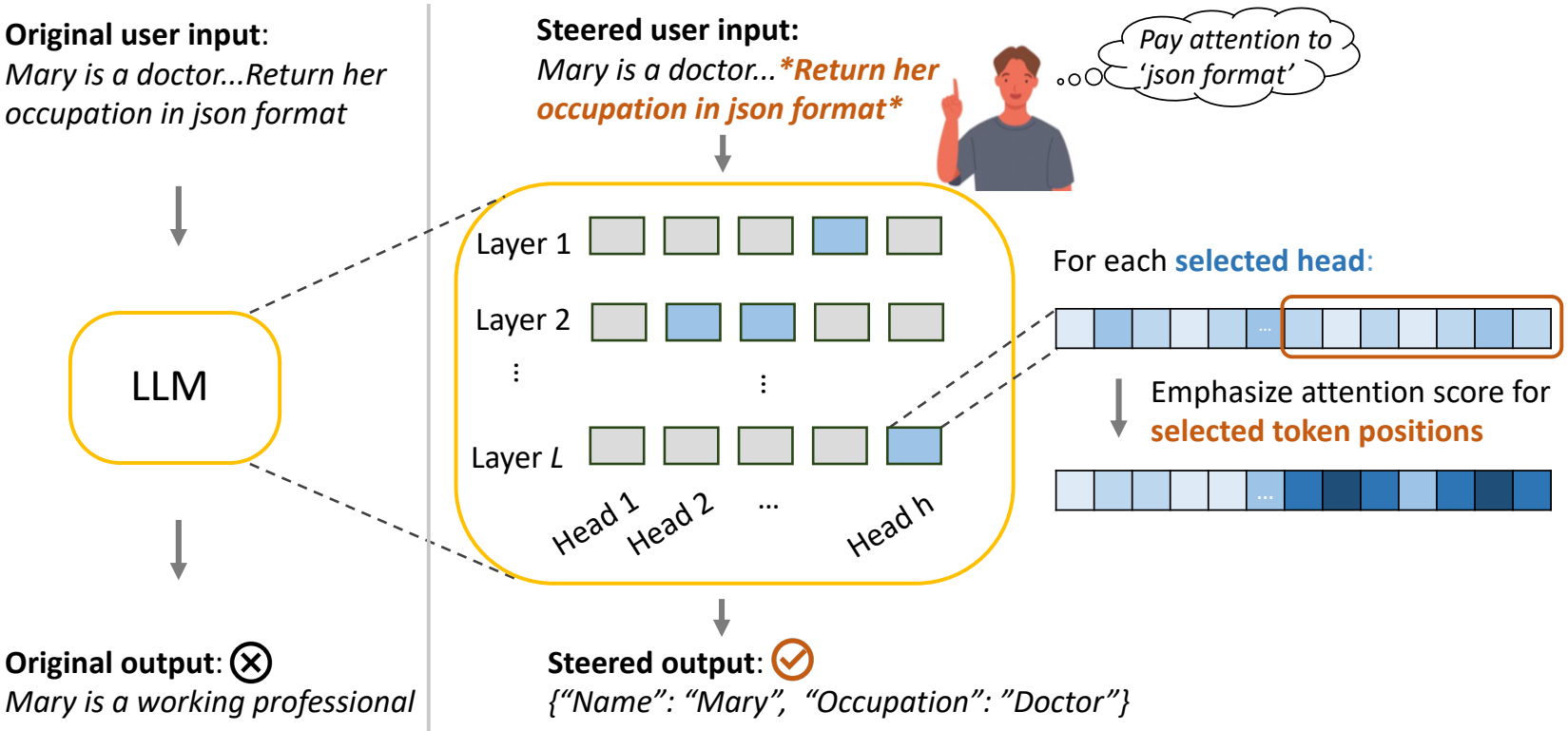
LLMs limitation

- LLMs are constrained to process plain-texts.
- **Question:** *Can we enable users to highlight specific texts and steer LLMs to interpret emphasized texts like human readers?*

Our Method: PASTA

PASTA: Post-hoc Attention STeering Approach

PASTA uses a user-specified part of the input to steer the model generation aligning with user intentions.



Post-hoc Attention Steering

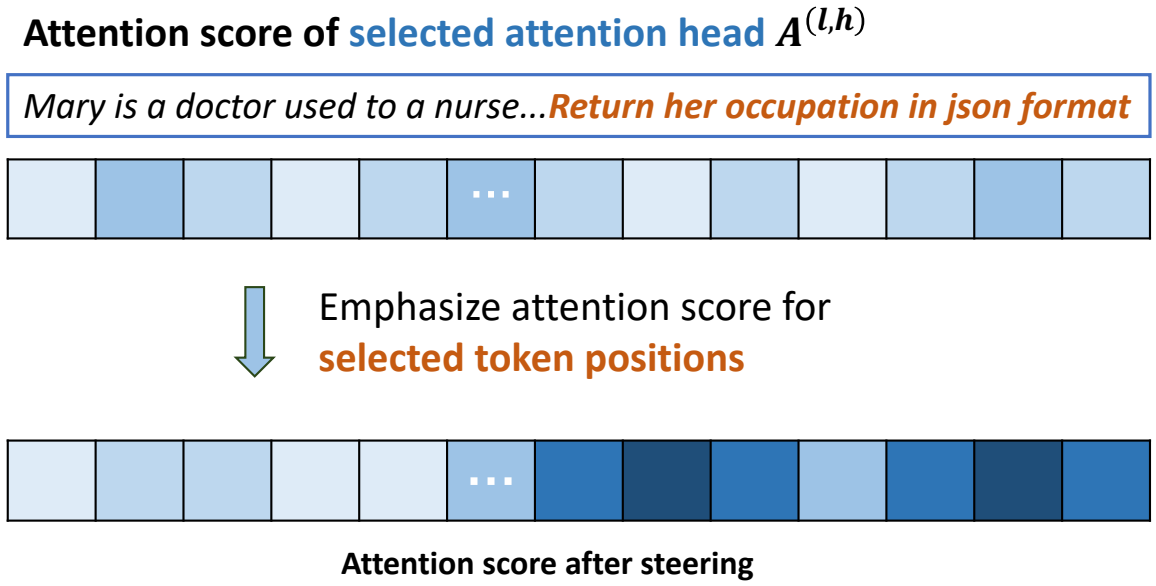
Steering models by attention reweighting

- Given the index set of highlighted input spans as $\mathcal{G}$,
- PASTA emphasizes these user-specified tokens by downweighting non-specified tokens ($\mathcal{G}^- = [n] - \mathcal{G}$):

$$\boldsymbol{H}^{(l,h)} = \mathcal{T}(\boldsymbol{A}^{(l,h)})\boldsymbol{V}, \quad [\mathcal{T}(\boldsymbol{A})]_{ij} = \begin{cases} \alpha \boldsymbol{A}_{ij}/C_i & \text{if } j \in \mathcal{G}^- \\ \boldsymbol{A}_{ij}/C_i & \text{o.w.} \end{cases}$$

- $0 < \alpha < 1$ is the scaling coefficient.
- $C_i = \sum_{j \in \mathcal{G}} \boldsymbol{A}_{ij} + \sum_{j \in \mathcal{G}^-} \alpha \boldsymbol{A}_{ij}$ normalizes the scores.

Post-hoc Attention Steering



- Attention scores of user-specified tokens are increased after renormalization.
- direct models to pay close attention to them.
- inference-only! No training at all.

Multi-Task Model Profiling

Steering performance varies dramatically across layers/heads

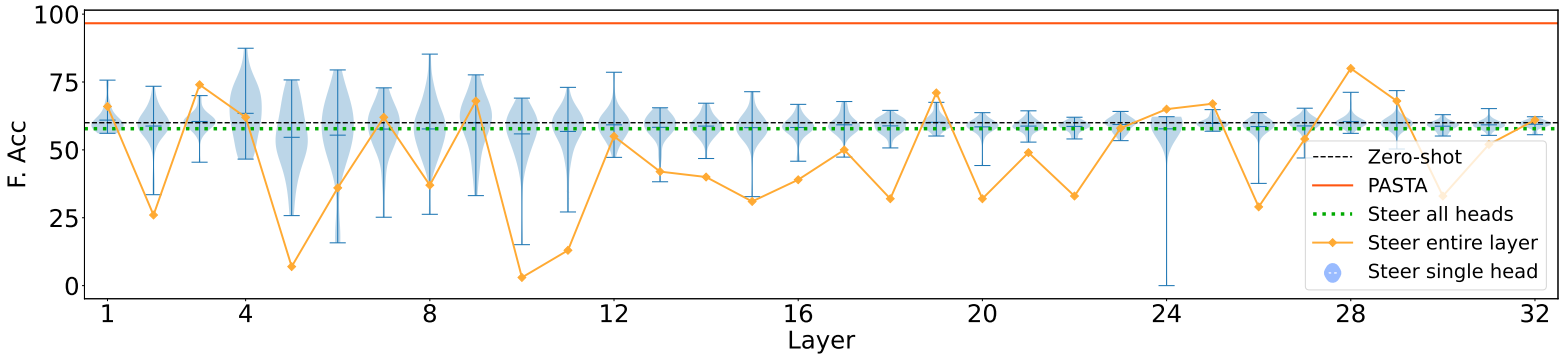


Figure: Performance of LLAMA-7B on JSON Formatting

- Different heads serve distinctive functions to encode semantic/syntactic information
- Some heads can be effectively steered across multiple tasks.
- Model profiling to identify effective heads.

Multi-Task Model Profiling

Model Profiling: Identify the effective heads

1. Subsample small training sets $\{\mathcal{D}^{(i)}\}_{i=1}^{m}$ (e.g., $|\mathcal{D}^{(i)}| = 1000$) from multiple tasks
2. Evaluate the performance of steering every head across tasks
3. Rank heads according to their performance on each task i: $R^{(i)} = [(l_1, h_1), (l_2, h_2), \dots]$
4. Select the head sets $\mathcal{H} = \cap_{i=1}^{m} R_{1:k}^{(i)}$.

Experiments

Evaluation Tasks

JSON Formatting

- Follow user instruction to return outputs in JSON.
- We emphasize the final instruction

Pronouns Changing

- Follow the instruction to change 'she/he' with 'they' in outputs.
- We emphasize the final instruction

BiasBios

- Predict occupation given lengthy contexts
- We emphasize the first sentence

CounterFact

- Answer the question given conflicting knowledge within contexts
- We emphasize the new facts in contexts.

Results of LLAMA-7B

Table: Main results of LLAMA-7B. For all scores, higher is better. The best results are in **bold**.

	Method	JSON Format F. Acc / P. Acc	Prons. Changing Acc / A.Acc	BiasBios Acc	CounterFact ES / PS	All Ave.
Prompting	Zero-shot	60.00 / 54.94	71.84 / 66.28	87.36	58.50 / 52.03	67.29
	$*$-marked	18.55 / 12.71	39.14 / 35.17	90.62	57.74 / 50.52	49.38
	“”-marked	4.56 / 4.20	20.55 / 18.19	89.82	58.14 / 51.70	42.15
	Few-shot	84.85 / 73.58	59.06 / 55.27	88.79	87.45 / 49.82	73.45
PASTA	Task-agnostic	88.16 / 49.08	83.65 / 81.31	93.54	98.82 / 99.03	85.89
	Multi-task	**96.64** / **85.09**	**96.42** / **95.84**	**95.28**	**99.60** / **99.57**	**95.46**

Results of GPT-J

Table: Main results of GPT-J. For all scores, higher is better. The best results are in **bold**.

	Method	JSON Format F. Acc / P. Acc	Prons. Changing Acc / A.Acc	BiasBios Acc	CounterFact ES / PS	All Ave.
Prompting	Zero-shot	28.83 / 25.09	39.88 / 36.19	72.76	42.14 / 42.02	44.96
	∗-marked	4.44 / 4.10	41.25 / 37.57	74.14	44.50 / 45.09	40.63
	“”-marked	8.81 / 5.62	6.12 / 5.72	78.64	45.54 / 41.84	33.87
	Few-shot	84.15 / **72.65**	35.77 / 32.08	72.98	68.34 / 38.23	59.65
PASTA	Task-agnostic	46.68 / 34.71	91.62 / 88.60	80.84	**99.54** / **99.57**	77.80
	Multi-task	**91.50** / 18.63	**92.96** / **91.34**	**94.96**	98.62 / 98.79	**85.22**

Thank You!