arXiv:2302.11520 (cs)

[Submitted on 22 Feb 2023 (v1), last revised 9 Oct 2023 (this version, v4)]

Guiding Large Language Models via Directional Stimulus Prompting

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, Xifeng Yan

We introduce Directional Stimulus Prompting, a novel framework for guiding black-box large language models (LLMs) toward specific desired outputs. Instead of directly adjusting LLMs, our method employs a small tunable policy model (e.g., T5) to generate an auxiliary directional stimulus prompt for each input instance. These directional stimulus prompts act as nuanced, instance-specific hints and clues to guide LLMs in generating desired outcomes, such as including specific keywords in the generated summary. Our approach sidesteps the challenges of direct LLM tuning by optimizing the policy model to explore directional stimulus prompts that align LLMs with desired behaviors. The policy model can be optimized through 1) supervised fine-tuning using labeled data and 2) reinforcement learning from offline or online rewards based on the LLM's output. We assess our method across summarization, dialogue response generation, and chain-of-thought reasoning tasks. Our experiments demonstrate that the framework consistently improves LLMs' (e.g., ChatGPT, Codex, InstructGPT) performance on these supervised tasks using minimal labeled data. Notably, using just 80 dialogues on the MultiWOZ dataset, our approach enhances ChatGPT's performance by an impressive 41.4%, matching or surpassing some fully supervised start-of-the-art models. Additionally, the instance-specific chain-of-thought prompt generated by our approach improves InstructGPT's reasoning accuracy compared to human-crafted or automatically generated prompts. The code and data are publicly available at \url{this https URL}.

Comments:	Accepted by NeurIPS2023. The code and data are available at this https URL
Subjects:	**Computation and Language (cs.CL)**
Cite as:	arXiv:2302.11520 **[cs.CL]**
	(or arXiv:2302.11520v4 **[cs.CL]** for this version)
	https://doi.org/10.48550/arXiv.2302.11520

Submission history

From: Zekun Li [view email]

[v1] Wed, 22 Feb 2023 17:44:15 UTC (1,650 KB)

[v2] Fri, 7 Apr 2023 18:19:03 UTC (2,303 KB)

[v3] Fri, 7 Jul 2023 01:31:49 UTC (699 KB)

[v4] Mon, 9 Oct 2023 21:01:22 UTC (3,915 KB)

Access Paper:

View PDF | TeX Source | Other Formats

Current browse context: **cs.CL**

< prev | next >

new recent 2023-02

Change to browse by: **cs**

References & Citations

NASA ADS Google Scholar Semantic Scholar

Export BibTeX Citation

Bookmark

Bibliographic Tools

Bibliographic and Citation Tools

○ Bibliographic Explorer (What is the Explorer?)

○ Litmaps (What is Litmaps?)

○ scite Smart Citations (What are Smart Citations?)

Code, Data, Media

Demos

Related Papers

About arXivLabs

Which authors of this paper are endorsers? | *Disable MathJax* (*What is MathJax?*)

About

Help

✉ Contact

Subscribe

Copyright

Privacy Policy

Web Accessibility Assistance

arXiv Operational Status

Get status notifications via email or slack