**arXiv:2306.12509** (cs)

[Submitted on 21 Jun 2023 (v1), last revised 4 Dec 2023 (this version, v2)]

# Joint Prompt Optimization of Stacked LLMs using Variational Inference

Alessandro Sordoni, Xingdi Yuan, Marc-Alexandre Côté, Matheus Pereira, Adam Trischler, Ziang Xiao, Arian Hosseini, Friederike Niedtner, Nicolas Le Roux

Large language models (LLMs) can be seen as atomic units of computation mapping sequences to a distribution over sequences. Thus, they can be seen as stochastic language layers in a language network, where the learnable parameters are the natural language prompts at each layer. By stacking two such layers and feeding the output of one layer to the next, we obtain a Deep Language Network (DLN). We first show how to effectively perform prompt optimization for a 1-Layer language network (DLN-1). Then, we present an extension that applies to 2-layer DLNs (DLN-2), where two prompts must be learned. The key idea is to consider the output of the first layer as a latent variable, which requires inference, and prompts to be learned as the parameters of the generative distribution. We first test the effectiveness of DLN-1 in multiple reasoning and natural language understanding tasks. Then, we show that DLN-2 can reach higher performance than a single layer, showing promise that we might reach comparable performance to GPT-4, even when each LLM in the network is smaller and less powerful.

Comments: NeurIPS 2023
Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)
Cite as: arXiv:2306.12509 **[cs.CL]**
(or arXiv:2306.12509v2 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2306.12509

## Submission history

From: Eric Yuan [view email]
[v1] Wed, 21 Jun 2023 18:45:56 UTC (206 KB)
**[v2]** Mon, 4 Dec 2023 15:07:13 UTC (212 KB)

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?*)

Get status notifications via email or slack