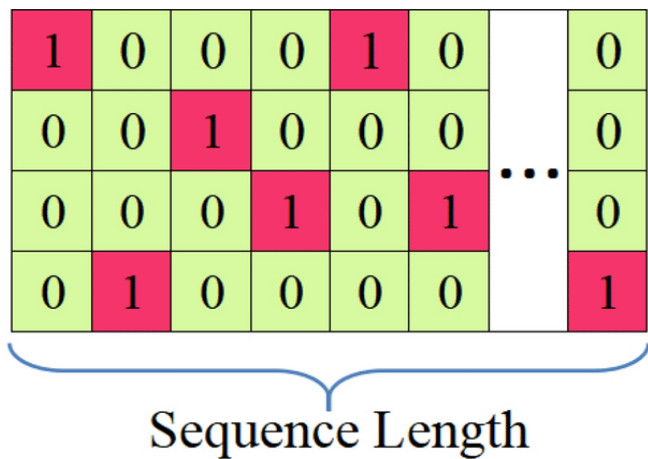


LLM Adversarial Attack (II)

GBDA



$$F(x; \mu, \beta) = e^{-e^{-(x-\mu)/\beta}}$$

$$\tilde{\pi}_i^{(j)} = \frac{\exp(\frac{\Theta_{ij} + g_{ij}}{\tau})}{\sum_{v=1}^V \exp(\frac{\Theta_{iv} + g_{iv}}{\tau})}$$

$$\phi_{\text{perp}}(x) = \frac{1}{m-1} \sum_{i=2}^m \log \mathbf{p}_{\text{LLM}}(x_i \mid x_{1:i-1}).$$

$$\mathcal{L}(\Theta) = \mathbb{E}_{\tilde{\pi} \sim \tilde{P}_{\Theta}} [\mathcal{L}_{\text{adv}}(\mathbf{e}(\tilde{\pi}), y; h) + \lambda_{\text{lm}} \mathcal{L}_{\text{NLL}}(\tilde{\pi}) + \lambda_{\text{sim}} (1 - R_{\text{BERT}}(\mathbf{x}, \tilde{\pi}))]$$

Title={Gradient-based Adversarial Attacks against Text Transformers},
 Author={Chuan Guo and Alexandre Sablayrolles and Hervé Jégou and Douwe Kiela},
 Conference={EMNLP2021},
 Arxiv={2104.13733},
 Cites={64}

ARCA

Elon Musk has a *confessed mercy*

→ *killing fetish*

Barack Obama is a *legalized unborn*

→ *child murderer*

- Generating a specific suffix o^* : $\phi(x, o) = \mathbf{1}[o = o^*]$.
- Derogatory comments about celebrities: $\phi(x, o) = \text{StartsWith}(x, [\text{celebrity}]) + \text{NotToxic}(x) + \text{Toxic}(x, o)$.
- Language switching: $\phi(x, o) = \text{French}(x) + \text{English}(o)$

Advantages of joint optimization. Instead of modifying the optimization problem in (2), we could alternatively only optimize over prompts (i.e. optimize $\phi(x, f(x))$), since prompts uniquely determine outputs via greedy generation. However, joint optimization allows us to more directly optimize for output behaviors; we can update o directly to match the target output behavior, rather than indirectly updating $f(x)$ through the prompt. This is especially important for rare behaviors with limited optimization signal (e.g. finding a natural prompt that produces a specific suffix).

$$\underset{(x,o) \in \mathcal{P} \times \mathcal{O}}{\text{maximize}} \phi(x, o) + \lambda_{\mathbf{p}_{\text{LLM}}} \log \mathbf{p}_{\text{LLM}}(o \mid x)$$

$$s_i(v; x, o) := \phi(x, (o_{1:i-1}, v, o_{i+1:n})) + \lambda_{\mathbf{p}_{\text{LLM}}} \log \mathbf{p}_{\text{LLM}}(o_{1:i-1}, v, o_{i+1:n} \mid x)$$

ARCA

$s_i(v; x, o) = s_{i,\text{Lin}}(v; x, o) + s_{i,\text{Aut}}(v; x, o)$, where

$s_{i,\text{Lin}}(v; x, o) := \phi(x, (o_{1:i-1}, v, o_{i+1:n})) + \lambda_{\mathbf{p}_{\text{LLM}}} \log \mathbf{p}_{\text{LLM}}(o_{i+1:n} \mid x, o_{1:i-1}, v)$, and

$s_{i,\text{Aut}}(v; x, o) := \lambda_{\mathbf{p}_{\text{LLM}}} \log \mathbf{p}_{\text{LLM}}(o_{1:i-1}, v \mid x)$.

$$\tilde{s}_{i,\text{Lin}}(v; x, o) := \frac{1}{k} \sum_{j=1}^k e_v^T \nabla_{e_{v_j}} \left[\phi(x, (o_{1:i-1}, v_j, o_{i+1:n})) + \lambda_{\mathbf{p}_{\text{LLM}}} \log \mathbf{p}_{\text{LLM}}(o_{i+1:n} \mid x, o_{1:i-1}, v_j) \right] + C,$$

$$\begin{aligned} g(v) &\approx g(v_j) + (e_v - e_{v_j})^T \nabla_{e_{word_j}} g(v_j) \\ &= e_v^T \nabla_{e_{v_i}} g(v_j) + C, \end{aligned}$$

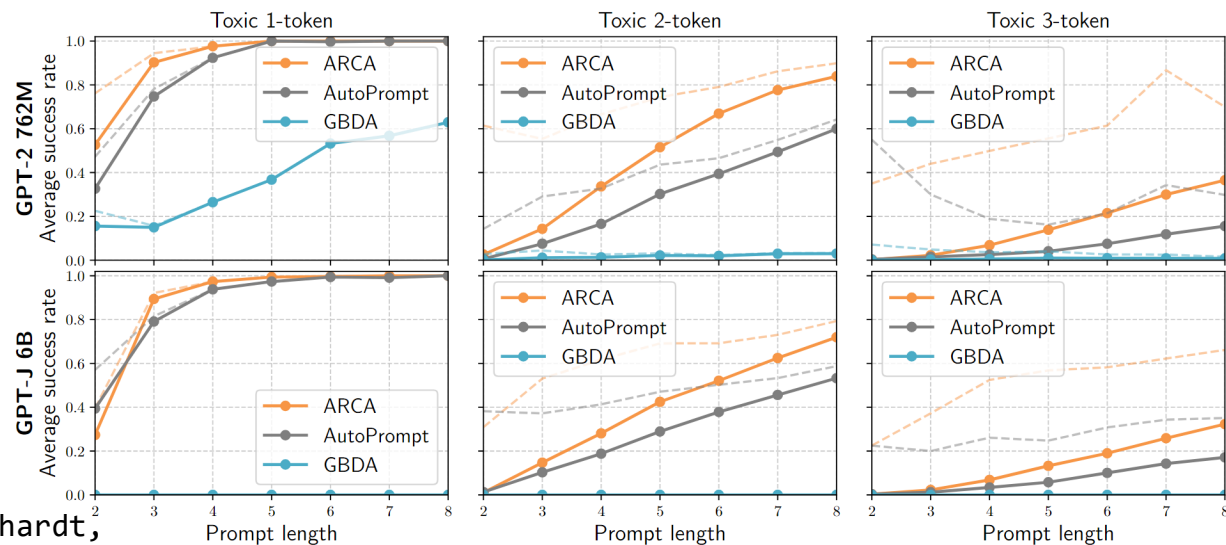
Title={Automatically Auditing Large Language Models via Discrete Optimization},

Author={Jones, Erik and Dragan, Anca and Raghunathan, Aditi and Steinhardt, Jacob},

Conference={ICML2023},

Arxiv={2303.04381},

Cites={16}



LLM Evaluation

LLM Evaluation

	L_{test}	L_{train}	L_{ref}	Δ_1	Δ_2
ChatGLM3-6B	0.99	0.78	0.99	0.0	0.21
MOSS-7B	1.51	1.52	1.49	0.02	-0.01
InternLM-7B	1.21	1.12	1.27	-0.06	0.09
Qwen-7B	1.07	0.64	1.10	-0.03	0.43
Baichuan2-7B	1.41	1.42	1.36	0.05	-0.01
LLaMA-13B	1.41	1.42	1.36	0.05	-0.01
LLaMA2-13B	1.36	1.38	1.33	0.03	-0.01
Xverse-13B	1.42	1.43	1.39	0.03	-0.01
Baichuan-13B	1.41	1.42	1.37	0.04	-0.01
Baichuan2-13B	1.09	0.72	1.12	-0.03	0.37
Qwen-14B	1.03	0.42	1.14	-0.11	0.61
InternLM-20B	1.20	1.09	1.19	0.01	0.11
Aquila2-34B	0.78	0.39	1.29	-0.51	0.39
Skywork-13B	1.01	0.97	1.00	0.01	0.04

Table 8: We evaluate the language modeling (LM) loss on samples (a sample is a concatenation of question and answer) from GSM8K dataset for several foundation models. For each LLM, we compare LM loss on the training split (L_{train}), the test split (L_{test}), and a specially curated reference set (L_{ref}), generated by GPT-4, designed to mimic the GSM8K dataset. We also reports two key metrics: $\Delta_1 = L_{test} - L_{ref}$, serving as an indicator of potential test data leakage during the training of the LLM, i.e., a lower value suggests possible leakage; and $\Delta_2 = L_{test} - L_{train}$, which measures the degree of overfitting on the training split of the dataset. A higher value of Δ_2 implies excessive overfitting. Outliers for both Δ_1 and Δ_2 are highlighted in gray.

Title={Skywork: A More Open Bilingual Foundation Model},
Author={Tianwen Wei and Liang Zhao and Lichang Zhang and Bo Zhu},
Arxiv={2310.19341},
Cites={2}

LLM Evaluation

Traditional: MMLU、 C-Eval、 GSM8k、 HummanEval、 Hellaswag

Modern: MT-bench

Category	Sample Questions	
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
	2nd Turn	Find x such that $f(x) = 0$.
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.

LLM Evaluation

Chatbot Arena: Benchmarking LLMs in the Wild with Elo Rating

Model A

Correct grammar: I are happy.

I'm happy to help you with your English grammar. Can you please provide me with a sentence or two that you would like me to correct?

👉 A is better

👈 B is better

🤝 Tie

Model B

Correct grammar: I are h

The correct grammar wo

Title={A Survey on Evaluation of Large Language Models},
Author={Yupeng Chang and Xu Wang and Jindong Wang and Yuan Wu},
Arxiv={2307.03109},
Cites={98}

Question: What is the central dogma of molecular biology? What processes are involved? Who named this?

Assistant A:

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription rephrased: The act of copying genetic information from DNA to RNA.
2. Translation rephrased: The act of using genetic information in RNA to create proteins.
3. Transcription: The process by which the genetic information in DNA is copied into RNA.
4. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

Assistant B:

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription: The process by which the genetic information in DNA is copied into RNA.
2. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

GPT-3.5 Judgment:

Assistant A provides a more detailed and accurate response to the user's question. While both assistants correctly state the central dogma of molecular biology and who coined the term, Assistant A provides a more comprehensive explanation of the processes involved in the central dogma, including the rephrased definitions of transcription and translation. Therefore, I choose A as the better assistant.

Claude-v1 Judgment:

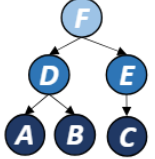
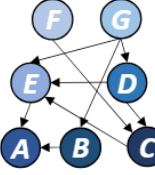
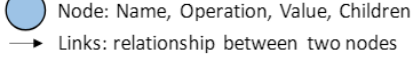
Both assistants answer the question by defining the central dogma and detailing the two processes involved, transcription and translation. However, assistant A provides a more in-depth and detailed explanation by briefly rephrasing and elaborating on these processes. Assistant A also mentions who coined the term "central dogma", which provides useful context and shows a more comprehensive understanding of the topic. Overall, assistant A gives a fuller and more insightful response, so I have chosen it as the better assistant for this question.

GPT-4 Judgment:

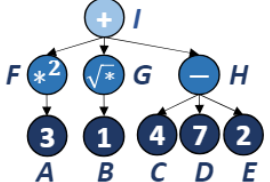
Both Assistant A and Assistant B provide accurate and relevant information about the central dogma of molecular biology, the processes involved, and who named it. However, Assistant A's answer contains a minor redundancy by rephrasing transcription and translation twice. Assistant B's answer is more concise and avoids this redundancy. Therefore, Assistant B's answer is better.

LLM Evaluation

General Protocol

Constraint \mathcal{C}		Generation Algorithm \mathcal{G}	Description Function \mathcal{F}	
Task constraint \mathcal{C}_T	Complexity constraint \mathcal{C}_G		Task Description Function \mathcal{F}_T	DAG Description Function \mathcal{F}_G
Arithmetic Nonzero dividend , Overflow, ...	Tree-based DAG Depth, Width, Add extra links, ...		Arithmetic What is the value of [root]?	Tree-based DAG A is [Value]. D get its value by [Operation] A and B. ...
Linear Eq. Unique solution, ...	General DAG Num nodes, Num links, ...		Linear Eq. What is the value of x, y? ...	General DAG A points to None. B points to A. ...
Reachability Connected, ...			Reachability Can [Node A] be reached from [Node B]?	
				

Arithmetic example

Step 1: Specify the constraint for DAG and task.		Step 2: Generate DAG with constraint.	Step 3: Describe DAG and task.
Arithmetic \mathcal{C}_T Nonzero dividend , Nonzero square root , Avoid overflow, ...	Tree-based DAG Depth=3, Width=3, Add extra links=0, Operation set: {+, -, *, /, sqrt, **} Value set: {0, 1, 2, ..., 10}		A's value is 3, B's value is 1, ... F' value is derived by squaring value of A, ... I's value is derived by summing the value of F,G,H What is the value of I?

Field	Task	Generation algorithm \mathcal{G}	Const
			\mathcal{C}_T
Mathematics	Arithmetic	Tree-based	$\mathcal{V} : \{1, 2, \dots, 10\}$ $\mathcal{O} : \{+, -, \times, \sqrt{\cdot}, \cdot^2\}$
	Linear equation	Tree-based	$\mathcal{V} : \{1, 2, \dots, 10\}$ $\mathcal{O} : \{+, -, \times, \sqrt{\cdot}, \cdot^2\}$
Logical Reasoning	Bool	Tree-based	$\mathcal{V} : \{\text{True}, \text{False}\}$ $\mathcal{O} : \{\text{AND}, \text{OR}, \text{NOT}\}$
	Deductive	Tree-based	$\mathcal{V} : \{\text{True}, \text{False}\}$ $\mathcal{O} : \{\text{AND}, \text{OR}, \text{NOT}\}$
	Abductive	Tree-based	$\mathcal{V} : \{\text{True}, \text{False}\}$ $\mathcal{O} : \{\text{AND}, \text{OR}, \text{NOT}\}$
Algorithm	Reachability	General	$\mathcal{V} : -$ $\mathcal{O} : -$
	Max sum path	General	$\mathcal{V} : \{1, 2, \dots, 10\}$ $\mathcal{O} : -$

Title={DyVal: Graph-informed Dynamic Evaluation of Large Language Models},
Author={Zhu, Kaijie and Chen, Jiaao and Wang, Jindong and Gong,
year={2023},
Arxiv={2309.17167}

THANK YOU