

ИГРА В ИМИТАЦИЮ:

О СПОСОБНОСТЯХ NLP-МОДЕЛЕЙ К РЕАЛЬНОМУ РАССУЖДЕНИЮ

**ПУТИЛОВА ЕЛЕНА,
ИССЛЕДОВАТЕЛЬ ДАННЫХ**

ТЕНДЕНЦИИ В NLP



- рост самих моделей



- появление prompt engineering

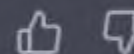
PROMPT: ОДИН ИЗ ПРИМЕРОВ



Напиши мне шутку про машинное обучение



Шутка про машинное обучение:



Почему машинное обучение так популярно?

Потому что оно никогда не спрашивает о повышении зарплаты, но всегда готово к работе!

PROMPTCAL: CONTRASTIVE AFFINITY LEARNING VIA AUXILIARY PROMPTS FOR GENERALIZED CATEGORY DISCOVERY

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh*
Hugging Face

Albert Webson*
Brown University

Colin Raffel*
Hugging Face

Stephen H. Bach
Brown & Snor

Lintang Sutawika
Ritsumeikan

Zaid Alyafeai
KUTIM

Antoine Chaffin
IRICA & IMATA

Arnaud Stiegler
Hugging Face

Tevan L.
Hugging Face

PROMPTSUM: PLANNING WITH MIXED PROMPTS FOR PARAMETER-EFFICIENT CONTROLLABLE ABSTRACT

Prompt Certified Machine Unlearning with Randomized Gradient Smoothing and Quantization

Zijie Zhang¹
zzz0092@aubu
tz

PROMPT INJECTION: PARAMETERIZATION OF FIXED INPUTS

S-Prompts Learning with Pre-trained Transformers: An Occam's Razor for Domain Incremental Learning

Vabin Wang^{1,2}, Zhiyuan Huang^{2†}, Yifan

Prompt-Augmented Li Scaling Beyond the Limit of Few-

Hyunsoo Cho[†], Hyuhng Joon Kim[†], Jun
Sang-goo Lee[†], Kang Min Yo

[†] Seoul National University, [‡]NAVER AI LAB, [§]NAVER
{johyunsoo, heyjoonkim, juny116, s
{kangmin.yoo, sang.woo.lee}@navercorp.c

Abstract

Through in-context learning (ICL), large-scale language models are effective few-shot learners without additional model fine-tuning. However, the ICL performance does not scale well with the number of available training samples as it is limited by the inherent input length constraint of the un-



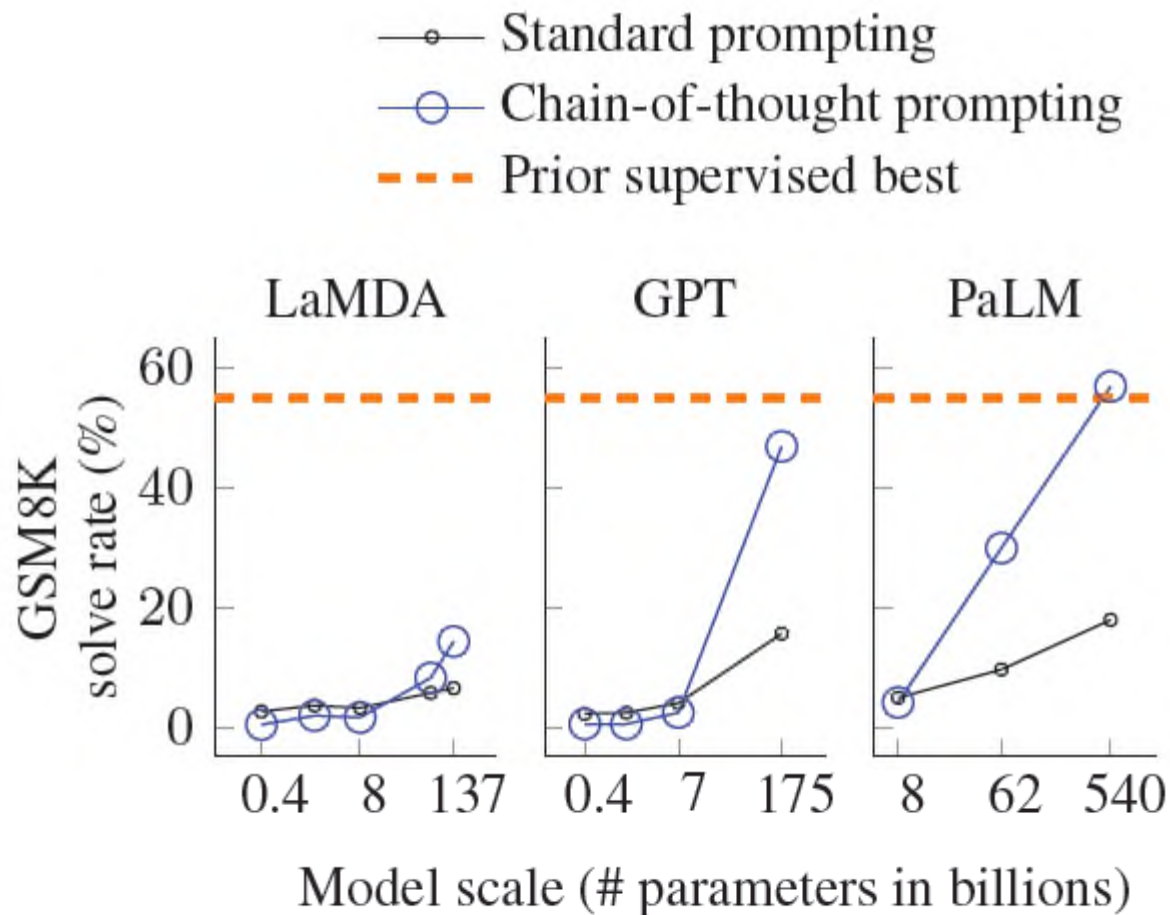
Ignore Previous Prompt: Attack Techniques For Language Models

Fábio Perez* Ian Ribeiro*
AE Studio
{fperez,ian.ribeiro}@ae.studio

Abstract

er-based large language models (LLMs) provide a powerful foundation language tasks in large-scale customer-facing applications. However, it explore their vulnerabilities emerging from malicious user interaction. By proposing PROMPTINJECT, a prosaic alignment framework based iterative adversarial prompt composition, we examine how GPT-4 widely deployed language model in production, can be easily mis-simply handcrafted inputs. In particular, we investigate two types of goal hijacking and prompt leaking – and demonstrate that even low-it sufficiently ill-intentioned agents, can easily exploit GPT-3's stochastic long-tail risks. The code for PROMPTINJECT is available at [om/agencyenterprise/PromptInject](https://github.com/agencyenterprise/PromptInject).

PROMPTING КАК «ХОД МЫСЛЕЙ»: РЕЗУЛЬТАТЫ*



- accuracy для PaLM540B на решении задач из GSM8K (датасет матем-х задач) возросло с 18% до 56,5%

*Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. **Chain of thought prompting elicits reasoning in large language models.** Conference on Neural Information Processing Systems (NeurIPS), 2022. URL <https://arxiv.org/pdf/2201.11903>

PROMPTING КАК «ХОД МЫСЛЕЙ»: КОРОТКО СУТЬ*

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

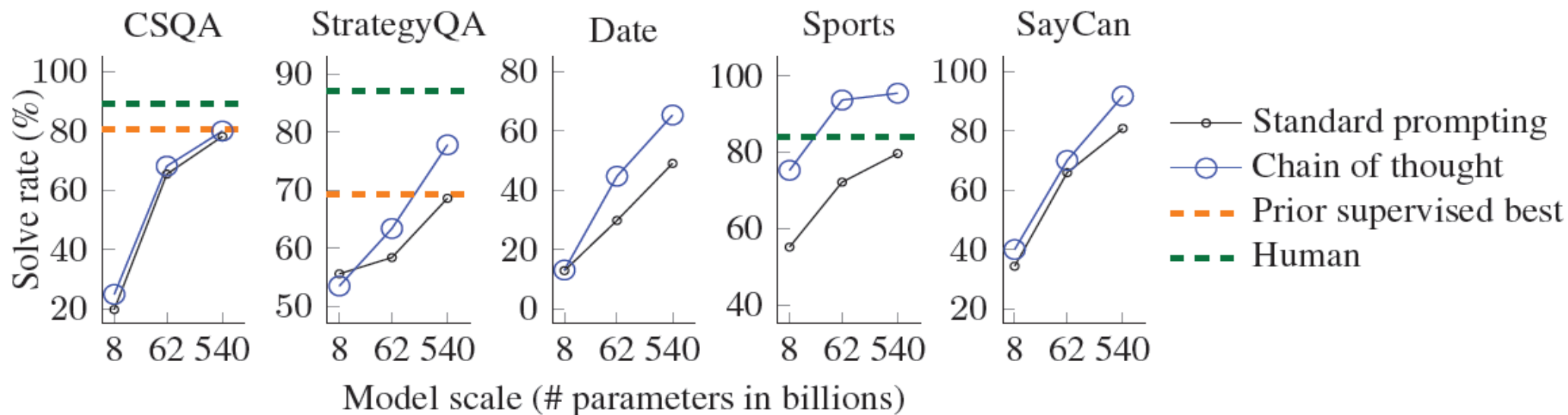
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

*Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. **Chain of thought prompting elicits reasoning in large language models.** Conference on Neural Information Processing Systems (NeurIPS), 2022. URL <https://arxiv.org/pdf/2201.11903>

МОДЕЛИ В ИССЛЕДОВАНИИ



PROMPTING КАК «ХОД МЫСЛЕЙ»: ЕЩЕ РЕЗУЛЬТАТЫ*



ПРИМЕР ДЛЯ ДАТАСЕТА SPORTS

QUESTION: Is the following sentence plausible? “Malcolm Brogdon eurostepped to the basket in the NBA Championship.”

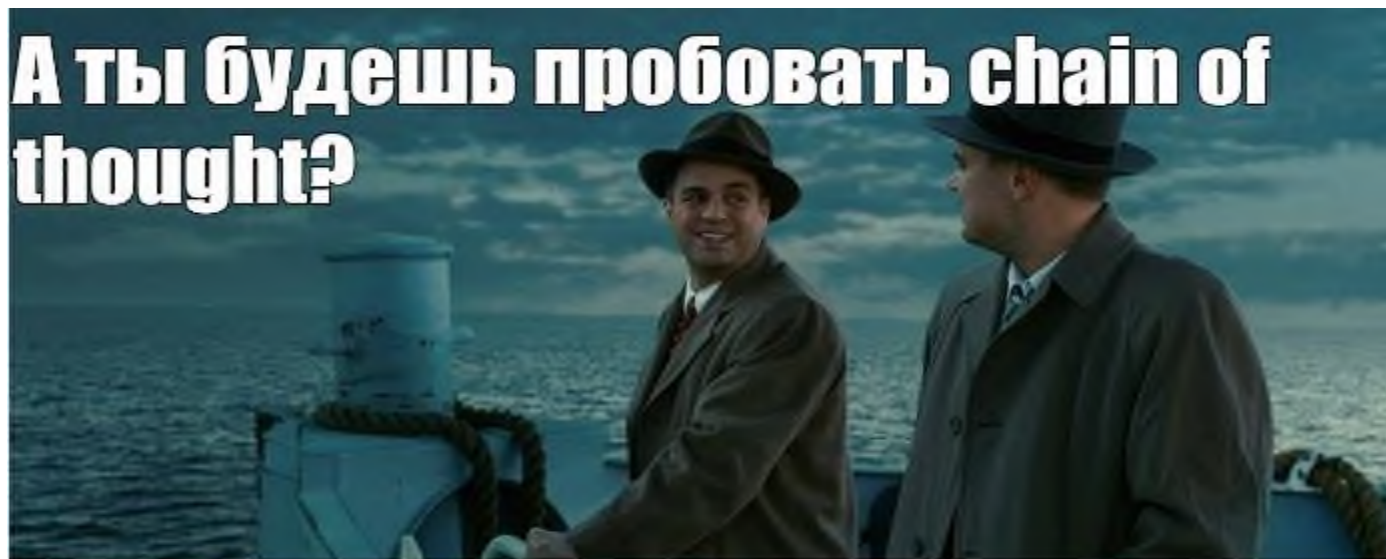
MODEL ANSWER (CORRECT): Malcolm Brogdon is a basketball player. Eurostepping to the basket is part of basketball. So the answer is yes. ✓

*Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou.

Chain of thought prompting elicits reasoning in large language models. Conference on Neural Information Processing Systems (NeurIPS), 2022. URL <https://arxiv.org/pdf/2201.11903>

ОСНОВНОЕ ОГРАНИЧЕНИЕ ПОДХОДА

А ты будешь пробовать chain of thought?



**Нет, у меня меньше 100В
параметров**



ВОПРОСЫ ДЛЯ ДАЛЬНЕЙШИХ ИССЛЕДОВАНИЙ

- Есть ли способ пробудить способности к «рассуждению» у «маленьких» моделей
- Как повлияет дальнейший рост количества параметров
- Как развитие prompt engineering может расширить диапазон задач

СПАСИБО ЗА ВНИМАНИЕ!