ИГРА В ИМИТАЦИЮ:

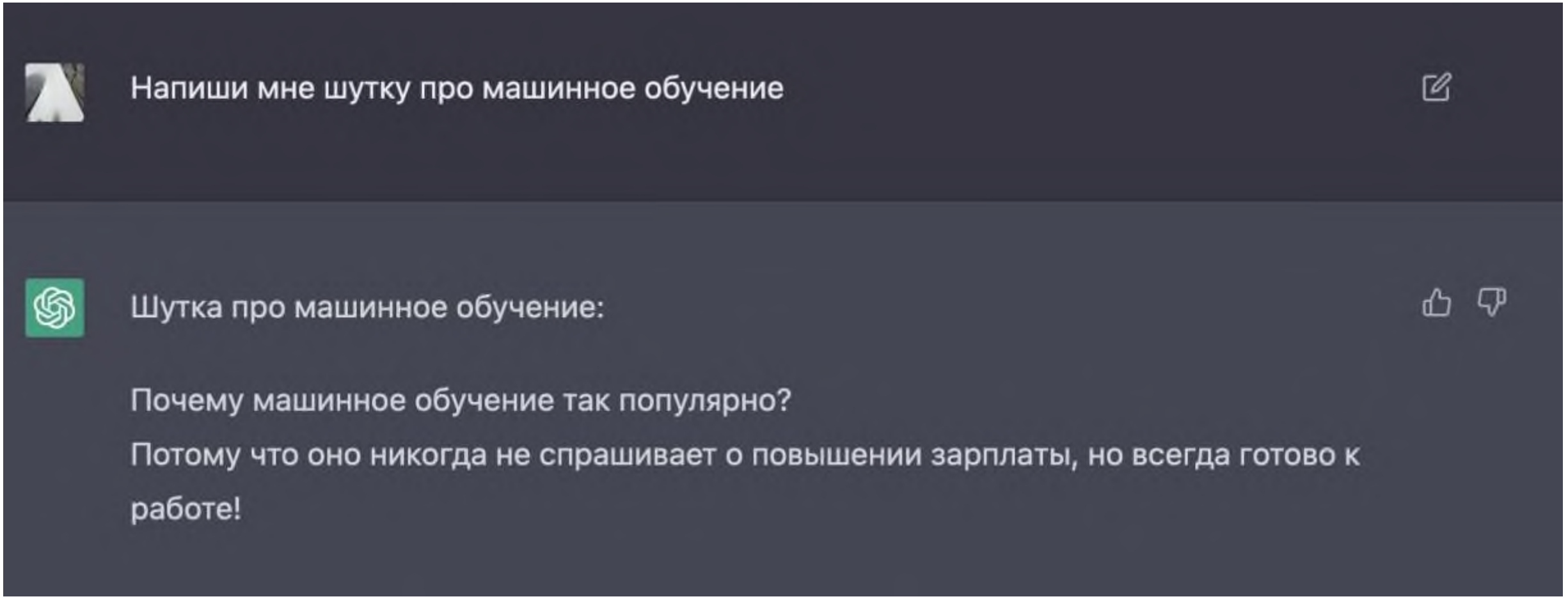
О СПОСОБНОСТЯХ NLP-МОДЕЛЕЙ К РЕАЛЬНОМУ РАССУЖДЕНИЮ

ПУТИЛОВА ЕЛЕНА,
ИССЛЕДОВАТЕЛЬ ДАННЫХ

ТЕНДЕНЦИИ В NLP

- рост самих моделей
- появление prompt engineering

PROMPT: ОДИН ИЗ ПРИМЕРОВ



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

PROMPT INJECTION: PARAMETERIZATION OF FIXED INPUTS

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

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Ba Brown & Snorl

Lintang Sutawika **Zaid Alyafeai** **Antoine Chaffin** **Arnaud Stiegler** **Teven L**

Prompt Certified Machine Unlearning with Randomized Gradient Smoothing and Quantization

Zijie Zhang[1]
zzz0092@aubu
tz

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

Yabin Wang[1,2]* Zhiwu Huang[2]† Xia

Ignore Previous Prompt: Attack Techniques For Language Models

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

Prompt-Augmented Li Scaling Beyond the Limit of Few-s

Hyunsoo Cho†, **Hyuhng Joon Kim**†, **Jun**
Sang-goo Lee†, **Kang Min Yo**
† Seoul National University, ‡NAVER AI LAB, §NAVE
{johyunsoo, heyjoonkim, juny116, s
{kangmin.yoo, sang.woo.lee}@navercorp.c

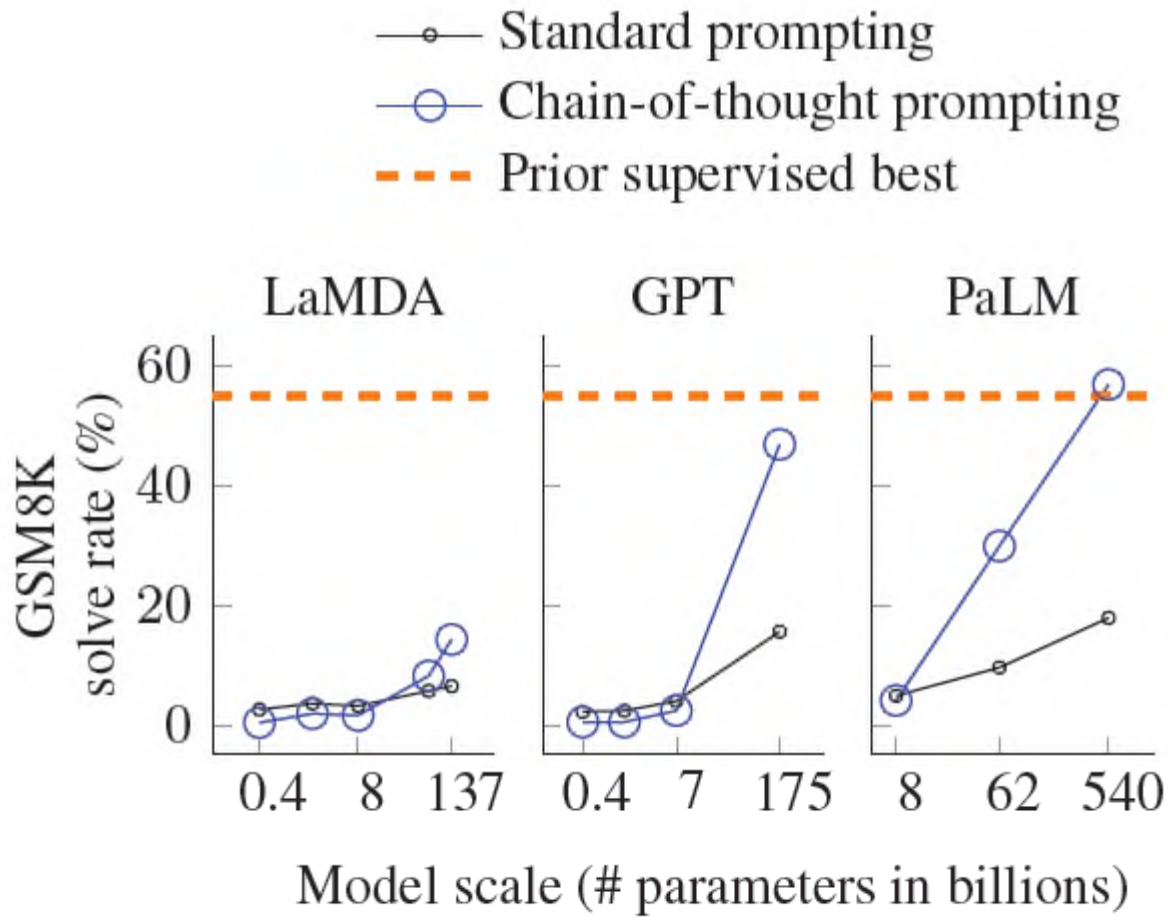
Abstract

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-

Abstract

er-based large language models (LLMs) provide a powerful foundation language tasks in large-scale customer-facing applications. However, t explore their vulnerabilities emerging from malicious user interac- arce. By proposing PROMPTINJECT, a prosaic alignment framework ased iterative adversarial prompt composition, we examine how GPT- t widely deployed language model in production, can be easily mis- simple handcrafted inputs. In particular, we investigate two types of oal hijacking and prompt leaking – and demonstrate that even low- t sufficiently ill-intentioned agents, can easily exploit GPT-3's stochas- creating long-tail risks. The code for PROMPTINJECT is available at om/agencyenterprise/PromptInject.

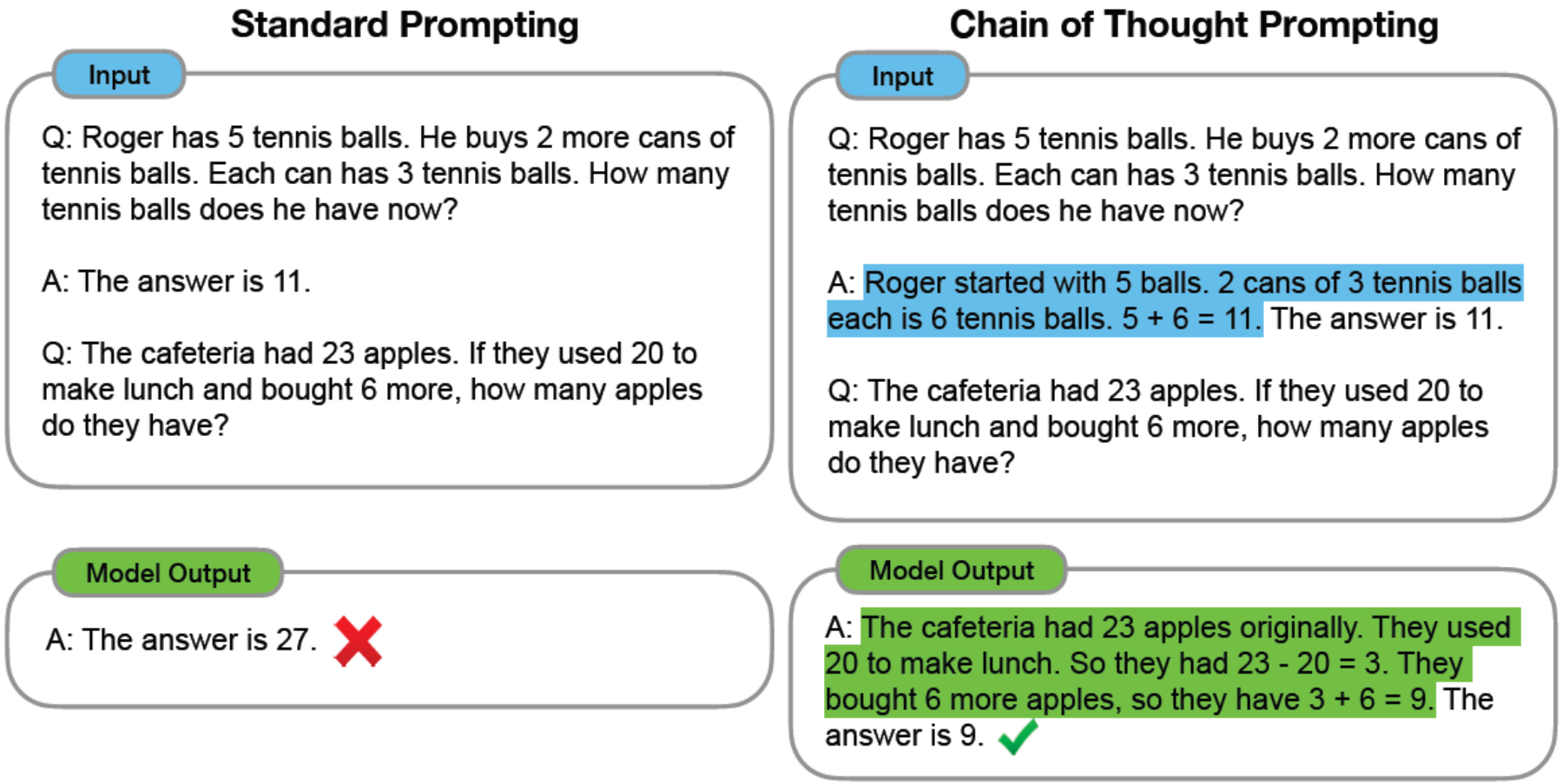
PROMPTING КАК «ХОД МЫСЛЕЙ»: РЕЗУЛЬТАТЫ*



- 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*

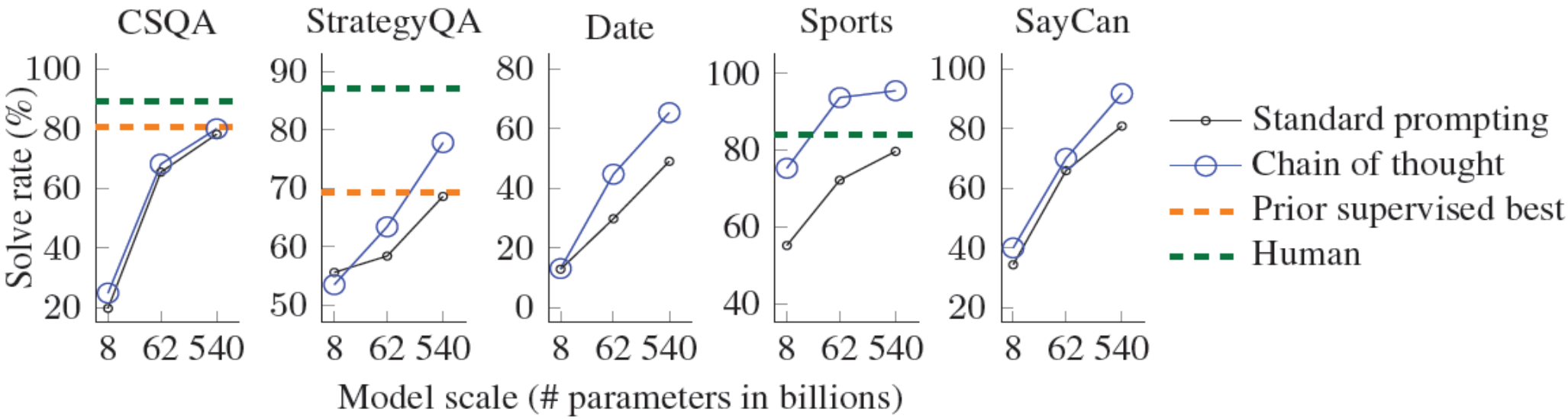
PROMPTING КАК «ХОД МЫСЛЕЙ»: КОРОТКО СУТЬ*



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

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!