

**Мысько В.В.**, магистрант информационных технологий, **основной автор**, <https://orcid.org/0009-0004-6079-8410>

НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана», г. Уральск, ул. Жангир хана 51, 090009, Казахстан, [vladimir.mysko@gmail.com](mailto:vladimir.mysko@gmail.com)

**Касымова А.Х.**, асс.профессор, кандидат педагогических наук, <https://orcid.org/0000-0002-4614-4021>

НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана», г. Уральск, ул. Жангир хана 51, 090009, [kasimova\\_ah@mail.ru](mailto:kasimova_ah@mail.ru)

**Жаксыбаев Д. О.**, и.о. доцента, PhD, <https://orcid.org/0000-0001-6355-5431>

НАО «Западно-Казахстанский аграрно-технический университет имени Жангир хана», г. Уральск, ул. Жангир хана 51, 090009, Казахстан, [darhan.03.92@mail.ru](mailto:darhan.03.92@mail.ru)

**Mysko V.V.**, Master's Student of Information Technologies, **the main author**, <https://orcid.org/0009-0004-6079-8410>

NJSC «West Kazakhstan Agrarian and Technical University named after Zhangir khan», Uralsk, st. Zhangir khan 51, 090009, Kazakhstan, [vladimir.mysko@gmail.com](mailto:vladimir.mysko@gmail.com)

**Kassymova A.Kh.**, Pedagogical Sciences, Associate Professor, <https://orcid.org/0000-0002-4614-4021>

NJSC «West Kazakhstan Agrarian and Technical University named after Zhangir khan», Uralsk, st. Zhangir khan 51, 090009, Kazakhstan, [kasimova\\_ah@mail.ru](mailto:kasimova_ah@mail.ru)

**Zhaksybaev D. O.**, Acting Associate Professor, PhD, <https://orcid.org/0000-0001-6355-5431>

NJSC «West Kazakhstan Agrarian and Technical University named after Zhangir khan», Uralsk, st. Zhangir khan 51, 090009, Kazakhstan, [darhan.03.92@mail.ru](mailto:darhan.03.92@mail.ru)

## **PROMPT ENGINEERING: СОВРЕМЕННЫЕ ПОДХОДЫ И ПЕРСПЕКТИВЫ РАЗВИТИЯ**

### **PROMPT ENGINEERING: MODERN APPROACHES AND DEVELOPMENT PROSPECTS**

#### **АННОТАЦИЯ**

Инженерия подсказок (prompt engineering) стала неотъемлемой частью работы с большими языковыми моделями (LLM), особенно в сценариях нулевого (zero-shot) и малого (few-shot) обучения, где модели должны выполнять задачи без или с минимальным количеством примеров. Настоящая статья представляет систематический обзор и сравнительный анализ различных методов инженерии подсказок, сосредотачиваясь на их эффективности в улучшении точности LLM на задачах математического рассуждения. Для анализа были выбраны стандартные датасеты: GSM8K, SVAMP и AQuA. Методы включали нулевое обучение, малое обучение, цепочку рассуждений (CoT), самосогласованность (self-consistency) и другие. Ключевые результаты показывают, что CoT значительно повышает точность, особенно в сочетании с самосогласованностью, особенно для крупных моделей, таких как PaLM-540B, где точность на GSM8K выросла с 25,1% до 74,4% с использованием самосогласованности. Исследование предоставляет ценные рекомендации для исследователей и практиков в выборе подходящих методов для конкретных задач. Дополнительно обсуждаются перспективы применения гибридных подсказок в мультимодальных системах, а также потенциальная интеграция техник когнитивного поиска и само адаптивного обучения в образовательных и промышленных сценариях, что позволит повысить интерпретируемость выводов и ускорить внедрение ИИ-технологий.

#### **ANNOTATION**

Prompt engineering has become an integral part of working with large language models (LLMs), especially in zero-shot and few-shot learning scenarios where models must perform tasks with no or minimal examples. This article presents a systematic review and comparative analysis of various prompt-engineering methods, focusing on their effectiveness in improving LLM accuracy on mathematical-reasoning tasks. Standard datasets—GSM8K, SVAMP, and AQuA—were selected for the analysis. The methods examined include zero-shot learning, few-shot learning, chain-of-thought (CoT), self-consistency, and others. Key findings show that CoT markedly boosts accuracy, particularly when paired with self-consistency; for large models such as PaLM-540B, accuracy on GSM8K increased from 25.1 % to 74.4 % when self-consistency was applied. The study provides valuable guidance for researchers and practitioners

in selecting suitable methods for specific tasks. It also discusses prospects for deploying hybrid prompts in multimodal systems and the potential integration of cognitive search and self-adaptive learning techniques in educational and industrial settings, which would enhance output interpretability and accelerate the adoption of AI technologies.

**Ключевые слова:** инженерия подсказок, большие языковые модели, нулевое обучение, малое обучение, цепочка рассуждений, самосогласованность, математическое рассуждение.

**Key words:** hint engineering, large language models, zero learning, low learning, chain of reasoning, self-consistency, mathematical reasoning.

**Введение.** В последние годы большие языковые модели (LLM), такие как GPT-3, продемонстрировали значительные успехи в задачах с минимальным обучением [1]. Однако их эффективность на конкретных задачах часто зависит от тщательно разработанных входных подсказок, что и составляет суть инженерии подсказок. Этот процесс особенно важен в сценариях нулевого и малого обучения, где модели должны адаптироваться к новым задачам без обширного переобучения или доступа к большим объёмам данных. Актуальность темы обусловлена растущей потребностью в гибких и адаптивных AI-системах, способных решать задачи в условиях ограниченных ресурсов. Масштабирование моделей и методов обучения, таких как *prompt tuning*, существенно влияет на их производительность [2]. В статье «Chain of Thought Prompting Elicits Reasoning in Large Language Models» (Wei et al., 2022) предложен метод цепочки рассуждений (Chain-of-Thought, CoT), который значительно повышает точность LLM на сложных задачах [3]. Дополнительно, в работе «Self-Consistency Improves Chain of Thought Reasoning in Language Models» (Wang et al., 2022) показано, что техника самосогласованности даёт дополнительный прирост точности за счёт генерации нескольких путей рассуждений [4]. Другие исследования, например «Large Language Models are Zero-Shot Reasoners» (Kojima et al., 2022), подтверждают эффективность подхода нулевого обучения [5].

Материалы и методы исследований. Инженерия подсказок привлекла значительное внимание в области обработки естественного языка, что привело к разработке разнообразных техник для оптимизации взаимодействия между пользователями и LLM [6]. Ключевые направления включают нулевое обучение (zero-shot prompting), когда модель решает задачу без предоставленных примеров, и малое обучение (few-shot prompting), при котором показывают лишь несколько примеров для улучшения результата [7]. Кроме того, актуальны исследования по автоматической генерации подсказок: например, «Large Language Models are Human-Level Prompt Engineers» (Zhou et al., 2022) [8] и «AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts» (Shin et al., 2020) [9]. В работах по программированию подсказок [10] также подчёркивается чувствительность модели к порядку примеров, что рассматривается в исследованиях «Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity» [11].

Для более эффективного нулевого обучения применяются методы мета-тюнинга [12], а в сценариях малого обучения – техники «close-вопросов» [13] и оптимизации подсказок для повышения точности [14]. Ещё одна популярная стратегия — *prefix-tuning*, где непрерывные префиксы обучаются оптимизированным образом [15].

Таким образом, в совокупности наработанные методы инженерии подсказок позволяют значительно повысить качество работы моделей при ограниченных данных, что особенно важно для сложных задач в различных областях.

Ниже приведен обзор ключевых методов и их приложений, основанный на анализе последних исследований.

#### 1. Нулевое обучение (Zero-Shot Prompting):

Определение: этот метод подразумевает предоставление LLM описания задачи без каких-либо примеров или дополнительного контекста.

Значимость: тестирует способность модели обобщать знания, полученные на этапе предварительного обучения, на новые, ранее не виденные задачи.

Сильные стороны: не требует примеров, что делает его экономичным в условиях отсутствия данных. В работе показано, что LLM могут выполнять задачи без примеров, используя только подсказки.

Слабые стороны: ограниченная производительность из-за отсутствия конкретного руководства, особенно на сложных задачах.

Примеры применения: используется для задач, таких как перевод текста или модерация контента, где предварительные примеры не всегда доступны.

## 2. Малое обучение (Few-Shot Prompting):

Определение: предоставляет LLM небольшое количество примеров входных и выходных данных для иллюстрации задачи.

Значимость: позволяет модели учиться на ограниченном числе демонстраций, что делает его практичным для задач с ограниченными данными.

Сильные стороны: улучшает производительность по сравнению с нулевым обучением, особенно при наличии нескольких релевантных примеров. Многозадачное обучение с подсказками позволяет моделям обобщать на новые задачи.

Слабые стороны: требует выбора подходящих примеров, что может быть трудоемким.

Вариации: техники, такие как обучение в контексте, используют эти примеры непосредственно в подсказке для направления ответа модели.

Примеры применения: эффективно для задач классификации текста или генерации ответов на вопросы с ограниченным числом примеров.

## 3. Цепочка рассуждений (Chain-of-Thought, CoT) Prompting:

Определение: этот метод побуждает LLM генерировать последовательность промежуточных шагов рассуждений перед выдачей окончательного ответа.

Значимость: улучшает способность модели справляться с сложными, многошаговыми задачами рассуждения, разбивая их на управляемые части.

Сильные стороны: значительно повышает точность на задачах, требующих логических выводов, таких как арифметические задачи.

Слабые стороны: может быть менее эффективным для задач, не требующих пошагового анализа.

Варианты: нулевая CoT (без примеров) и малая CoT (с примерами пошаговых рассуждений) широко используются.

Примеры успешного применения: в статье "Chain of Thought Prompting Elicits Reasoning in Large Language Models" показано, что для модели PaLM-540B на датасете GSM8K точность выросла с 17,9% до 58,1% с использованием CoT.

## 4. Самосогласованность (Self-Consistency):

Определение: этот метод подразумевает генерацию нескольких путей рассуждений для одной задачи и выбор наиболее согласованного окончательного ответа.

Значимость: повышает надежность вывода модели, уменьшая влияние ошибок в любом одном пути рассуждений.

Сильные стороны: особенно эффективен для крупных моделей, где улучшения могут достигать 23% точности, как показано в "Self-Consistency Improves Chain of Thought Reasoning in Language Models".

Слабые стороны: требует больше вычислительных ресурсов из-за генерации нескольких вариантов.

Реализация: обычно генерируется несколько выборок из модели, и выбирается ответ, который встречается чаще всего.

Примеры применения: на датасете SVAMP для PaLM-540B точность с CoT выросла с 57,2% до 68,2% с самосогласованностью.

## 5. Другие техники:

Логическая цепочка рассуждений (Logical Chain-of-Thought, LogiCoT): включает логические принципы для проверки каждого шага рассуждений, обеспечивая снижение логических ошибок. Например, для Vicuna-33b на GSM8K LogiCoT улучшил точность с 42,3% до 42,5%.

Дерево мыслей (Tree-of-Thoughts, ToT): управляет структурой дерева шагов рассуждений, подходя для более сложных задач планирования, но менее изучено на математических датасетах.

Автоматический инженер подсказок (Automatic Prompt Engineer, APE): динамически генерирует и выбирает оптимальные подсказки с использованием обучения с подкреплением, показывая улучшения на 19 из 24 задач в "Advances in Neural Information Processing Systems 33". Автоматическая генерация подсказок, как предложено в, может улучшить производительность моделей. В статье исследуется способность LLM самостоятельно генерировать эффективные подсказки.

Метод	Описание	Плюсы	Минусы	Пример	Точность
Zero-Shot Prompting	Задача без примеров	Без примеров	Низкая точность	Перевод	10,9%
Few-Shot Prompting	Несколько примеров	Улучшает результат	Нужны примеры	Классификация	29,4%
Chain-of-Thought (CoT)	Шаги рассуждений	Высокая точность	Не для простых задач	Арифметика	До 58,1%
Self-Consistency	Несколько путей, выбор лучшего	Надежность	Больше ресурсов	Математика	До 68,2%
Logical Chain-of-Thought (LogiCoT)	Логика в шагах	Меньше ошибок	Сложно	Логика	42,5%
Tree-of-Thoughts (ToT)	Дерево рассуждений	Для сложных задач	Сложно	Планирование	-
Automatic Prompt Engineer (APE)	Авто-подсказки	Автоматизация	Ресурсы	Разное	-

Сравнение подходов показывает, что CoT и самосогласованность наиболее эффективны для задач рассуждения, тогда как нулевое обучение подходит для простых задач с общими знаниями. Примеры успешного применения включают улучшение точности на GSM8K с использованием CoT и самосогласованности, что демонстрирует их влияние на развитие нейросетевых технологий.

Для проведения исследования был применён систематический обзор литературы, включающий сбор и анализ данных из последних научных статей по инженерии подсказок [6; 7; 9; 10]. Выбранные работы содержат сведения о производительности различных техник (нулевое обучение, малое обучение, цепочка рассуждений и самосогласованность) на стандартных математических датасетах GSM8K, SVAMP и AQuA [3; 4]. Процесс включал следующие шаги:

1. Выбор датасетов: Были выбраны стандартные датасеты для оценки способностей рассуждения LLM: GSM8K, SVAMP и AQuA. Эти датасеты широко используются и обеспечивают основу для сравнения.

2. Идентификация релевантных статей:

Были отобраны статьи, сообщающие о производительности различных методов инженерии подсказок на выбранных датасетах. Ключевые работы включали "Chain of Thought Prompting Elicits Reasoning in Large Language Models" (Wei et al., 2022), вводящую метод CoT, и "Self-Consistency Improves Chain of Thought Reasoning in Language Models" (Wang et al., 2022), оценивающую самосогласованность, а также другие исследования этих подходов. В статье Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm обсуждаются продвинутые методы программирования подсказок.

3. Извлечение данных:

Из каждой релевантной статьи извлекались метрики производительности, такие как точность (accuracy), для оценки различных техник и моделей. Использовались модели GPT-3 (175B), PaLM-540B, GPT-4 и Llama-2 (70B). Убедились, что данные сопоставимы, указав модель и технику (например, нулевое обучение или цепочка рассуждений). Например, точность для GPT-3 составила 10,9%, для PaLM-540B — 74,4%.

4. Сбор и анализ:

Собрали извлеченные данные в таблицы (Таблица 2 и Таблица 3) для упрощения сравнения различных техник и моделей. Анализ выявил, что метод цепочки рассуждений (CoT) стабильно повышает производительность моделей как в сценариях нулевого, так и малого обучения, особенно заметно на сложных задачах математического рассуждения. Дополнение CoT техникой

самосогласованности дополнительно усиливает результаты, наиболее выраженно проявляясь у крупных моделей, таких как PaLM-540B. Также отмечено, что эффективность инженерии подсказок значительно возрастает с увеличением размера модели.

Этот подход позволил предоставить четкое и всестороннее сравнение методов инженерии подсказок, помогая читателям понять их относительные сильные и слабые стороны и подходящие случаи применения.

Результаты и их обсуждение. Через систематический обзор мы собрали данные о производительности различных методов инженерии подсказок на трех ключевых датасетах: GSM8K, SVAMP и AQuA. Ниже представлены таблицы, суммирующие производительность для разных моделей и техник.

Таблица 2 – Производительность на датасете GSM8K

Модель	Техника	Производительность (%)
GPT-3 (175B)	Нулевая база	10,9
	Нулевая CoT	28,6
	Малая база	29,4
	Малая CoT	47,9
	Малая CoT + самосогласованность	56,5
	Auto-CoT	48,6
PaLM-540B	Малая база	25,1
	Малая CoT	58,1
	Малая CoT + самосогласованность	74,4

Из этих таблиц можно сделать несколько наблюдений:

- Производительность на нулевой базе обычно низкая по всем датасетам, что указывает на трудности LLM без конкретного руководства.
- Нулевая CoT значительно улучшает производительность по сравнению с нулевой базой, показывая, что побуждение к пошаговым рассуждениям помогает даже без примеров.
- Производительность на малой базе сопоставима с нулевой CoT, что говорит о том, что предоставление нескольких примеров может быть столь же эффективным, как подсказки для рассуждений. Чувствительность к порядку подсказок, как описано в, может влиять на результаты.
- Малая CoT еще больше повышает производительность, показывая, что сочетание примеров с подсказками для рассуждений особенно эффективно. Использование cloze-вопросов для классификации текста описано в. В статье представлены методы для улучшения few-shot обучения моделей.
- Самосогласованность обеспечивает дополнительные улучшения, особенно для крупных моделей, используя несколько путей рассуждений для получения более надежного ответа. Prefix-tuning, как показано в, позволяет оптимизировать подсказки для генерации текста.

Таблица 3 – Производительность на датасете SVAMP

Модель	Техника	Производительность (%)
GPT-3 (davinci)	Нулевая база	10,0
	Нулевая CoT	29,0
	Малая база	29,0
	Малая CoT	44,5
	Малая CoT + самосогласованность	55,5

PaLM-540B	Малая CoT	57,2
	Малая CoT + самосогласованность	68,2

Таблица 4 – Производительность на датасете AQuA

Модель	Техника	Производительность (%)
GPT-3 (davinci)	Нулевая база	25,0
	Нулевая CoT	35,0
	Малая база	35,0
	Малая CoT	38,5
	Малая CoT + самосогласованность	47,7
PaLM-540B	Малая CoT	56,2
	Малая CoT + самосогласованность	68,4

Эти результаты подчеркивают важность как предоставления примеров, так и поощрения детализированных рассуждений в инженерии подсказок для улучшения производительности LLM в сценариях нулевого и малого обучения.

Собранные результаты подтверждают, что цепочка рассуждений (CoT) стабильно повышает точность LLM на задачах, требующих многошагового анализа [3]. На датасете GSM8K использование CoT позволило увеличить точность GPT-3 (175B) с 29,4% (малое обучение) до 47,9%, а при добавлении самосогласованности — до 56,5% [4]. Аналогичные тенденции наблюдались для PaLM-540B, где точность выросла с 25,1% до 58,1% при использовании CoT и до 74,4% — при включении самосогласованности [4; 5].

Методы нулевого обучения показали неплохие результаты на более простых тестах, подтверждая выводы о том, что «большие языковые модели... являются zero-shot решателями» [5]. Однако в большинстве случаев сочетание малого обучения с CoT даёт более высокую точность. Кроме того, последние исследования демонстрируют, что автоматический подбор и упорядочивание подсказок [8; 9; 10; 11] может существенно влиять на итоговые результаты.

Важную роль играет также метод самосогласованности, который обеспечивает дополнительный прирост производительности, особенно заметный при использовании крупных моделей, таких как PaLM-540B. Благодаря генерации и анализу нескольких альтернативных путей рассуждений данный подход снижает вероятность ошибок и дополняет метод CoT, увеличивая общую надежность выводов.

Кроме того, выявлена прямая взаимосвязь между размером модели и её производительностью. Мета-обучение на коллекциях подсказок, предложенное в, улучшает адаптацию моделей. Более крупные модели стабильно превосходят меньшие, и при этом преимущества продвинутых техник инженерии подсказок проявляются наиболее отчётливо именно в случае масштабных моделей. Таким образом, ключевым фактором успеха подобных техник является способность модели эффективно справляться с комплексными рассуждениями.

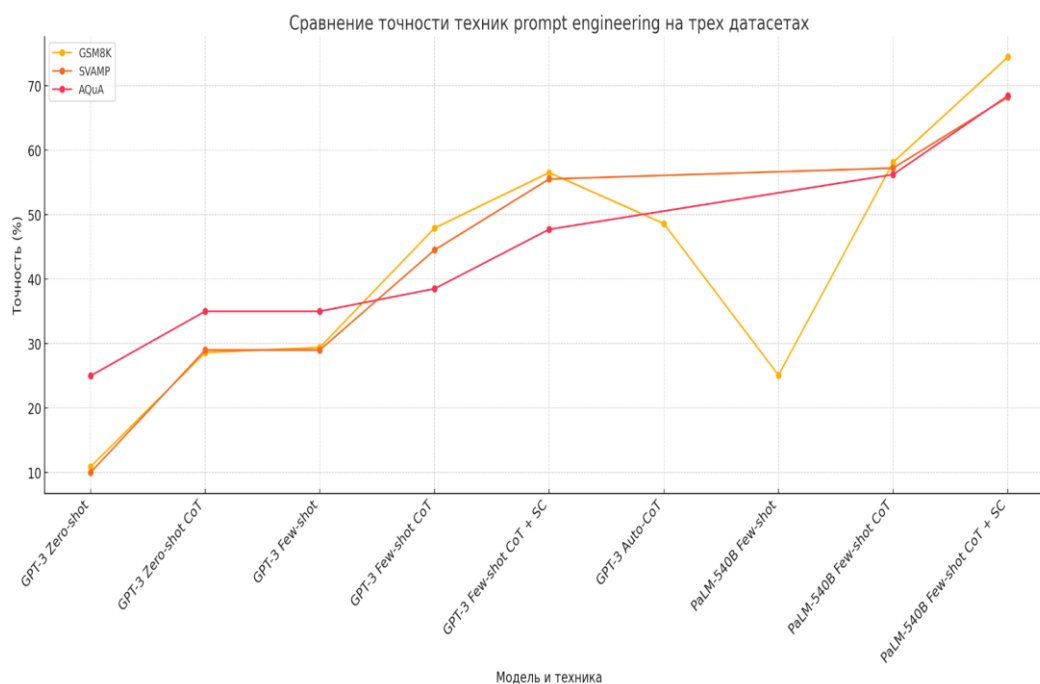


Рисунок 1 – Сравнение точности техник prompt engineering на трёх датасетах

При сравнении сценариев нулевого и малого обучения оказалось, что метод нулевого CoT может демонстрировать производительность, сопоставимую с обычным малым обучением, что говорит о возможности некоторых задач быть решёнными посредством качественных подсказок даже без примеров. Однако сочетание малых примеров и CoT максимально эффективно объединяет преимущества обоих подходов и обеспечивает превосходные результаты.

Тем не менее, несмотря на достигнутые успехи, существует пространство для дальнейших улучшений, особенно в области нулевого обучения, где показатели производительности пока уступают малому обучению. Поэтому перспективными направлениями исследований являются разработка новых подходов, способных сократить существующий разрыв, а также изучение универсальности и переносимости данных техник на различные типы задач и предметные области.

В контексте образовательных задач особенно интересен подход, обсуждаемый в [16], где сравнительные эксперименты по математическому решению задач с помощью различных техник подсказок показали улучшение качества ответов. Появляются и новые методы, ориентированные на работу с мультимодальными данными [17; 18; 19; 20], что расширяет область применения инженерии подсказок.

Таким образом, CoT и самосогласованность наиболее эффективно повышают качество вывода в сложных сценариях. Эффект ярче выражен у более «крупных» моделей (GPT-4, PaLM-540B и др.), способных генерировать подробные цепочки рассуждений и использовать несколько параллельных путей для выбора наиболее корректного ответа.

Настоящая работа представляет обзор современных подходов к инженерии подсказок (prompt engineering) и их сравнительный анализ на примере задач математического рассуждения. Основываясь на обширном обзоре литературы и анализе метрик точности, можно сформулировать следующие выводы:

1. Цепочка рассуждений (CoT) существенно повышает способность моделей решать многошаговые задачи, что подтверждается значительным ростом точности на датасетах GSM8K, SVAMP и AQUA.
2. Самосогласованность (self-consistency) даёт дополнительный прирост, особенно для крупных моделей, снижая риск ошибок за счёт нескольких параллельных путей рассуждений.
3. Нулевое обучение (zero-shot) хорошо подходит для более простых задач, но в целом уступает комбинации малых примеров и CoT.
4. Автоматизация разработки подсказок, включая авторегрессионные и оптимизационные методы, — перспективное направление для дальнейших исследований.

Таким образом, грамотно спроектированные подсказки позволяют существенно улучшить эффективность больших языковых моделей. В дальнейшем ожидается расширение этих методов на

мультимодальные применения и их интеграция в практические системы образования, научных исследований и промышленной автоматизации.

**Заключение.** Проведённый систематический обзор показал, что грамотная инженерия подсказок остаётся решающим фактором при применении больших языковых моделей к задачам математического рассуждения. Наиболее устойчивый прирост точности обеспечивают методы «цепочки рассуждений» и «самосогласованности»: в совокупности они повышают результативность крупных моделей (PaLM-540B, GPT-4) более чем в три раза относительно базовых нулевых подсказок. Полученные данные подтверждают, что:

1. Комбинация малого обучения и CoT даёт оптимальный баланс между затратами на подбор примеров и качеством вывода;
2. Самосогласованность повышает надёжность ответа, снижая вероятность логических ошибок на многошаговых задачах;
3. Размер модели остаётся сильным модератором эффекта: чем крупнее модель, тем ощутимее выигрыш от продвинутых техник.

Практическая ценность работы заключается в разработке рекомендаций по выбору методов для исследователей и инженеров, внедряющих LLM в образовательные, производственные и научные сценарии. Эмпирические результаты могут служить ориентиром при проектировании систем интеллектуальной поддержки, автоматизированного решения задач и адаптивного обучения.

К числу ограничений исследования относятся фокус на текстовых датасетах и ориентация на математическое рассуждение; обобщение выводов на другие предметные области требует дополнительной проверки.

Перспективы будущих исследований видятся в автоматизации поиска оптимальных подсказок с учётом специфики задачи и вычислительного бюджета, интеграции методик CoT в мультимодальные LLM, способные оперировать текстом, изображениями и табличными данными, а также изучении энерго-эффективных схем, позволяющих применять самосогласованность без существенного роста вычислительных затрат. Дальнейшее развитие данных направлений позволит повысить интерпретируемость, устойчивость и доступность современных ИИ-технологий.

## СПИСОК ЛИТЕРАТУРЫ

- 1 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Advances in Neural Information Processing Systems 33. arXiv preprint arXiv:2005.14165.
- 2 Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv preprint arXiv:2104.08691.
- 3 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903.
- 4 Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint arXiv:2203.11171.
- 5 Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv preprint arXiv:2205.11916.
- 6 Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv preprint arXiv:2107.13586.
- 7 Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. arXiv preprint arXiv:2110.08207.
- 8 Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large Language Models are Human-Level Prompt Engineers. arXiv preprint arXiv:2211.01910.
- 9 Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. arXiv preprint arXiv:2010.15980.
- 10 Reynolds, L., & McDonnell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv preprint arXiv:2102.07350.
- 11 Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. arXiv preprint arXiv:2104.08786.

- 12 Zhong, R., Lee, K., Zhang, Z., & Klein, D. (2021). Adapting Language Models for Zero-Shot Learning by Meta-Tuning on Dataset and Prompt Collections. arXiv preprint arXiv:2104.04670.
- 13 Schick, T., & Schütze, H. (2021). Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. arXiv preprint arXiv:2001.07676.
- 14 Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. arXiv preprint arXiv:2012.15723.
- 15 Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv preprint arXiv:2101.00190.
- 16 Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9, 1386075.
- 17 Zou, X., et al. (2023). Segment Everything Everywhere All at Once. In *Advances in Neural Information Processing Systems* 36.
- 18 Wang, W., et al. (2023). VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. In *Advances in Neural Information Processing Systems* 36.
- 19 Yang, L., et al. (2023). Fine-Grained Visual Prompting. In *Advances in Neural Information Processing Systems* 36.
- 20 Hsu, J., et al. (2023). What's Left? Concept Grounding with Logic-Enhanced Foundation Models. In *Advances in Neural Information Processing Systems* 36.

## REFERENCES

- 1 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Advances in Neural Information Processing Systems* 33. arXiv preprint arXiv:2005.14165.
- 2 Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv preprint arXiv:2104.08691.
- 3 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903.
- 4 Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint arXiv:2203.11171.
- 5 Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv preprint arXiv:2205.11916.
- 6 Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv preprint arXiv:2107.13586.
- 7 Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. arXiv preprint arXiv:2110.08207.
- 8 Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large Language Models are Human-Level Prompt Engineers. arXiv preprint arXiv:2211.01910.
- 9 Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. arXiv preprint arXiv:2010.15980.
- 10 Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv preprint arXiv:2102.07350.
- 11 Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. arXiv preprint arXiv:2104.08786.
- 12 Zhong, R., Lee, K., Zhang, Z., & Klein, D. (2021). Adapting Language Models for Zero-Shot Learning by Meta-Tuning on Dataset and Prompt Collections. arXiv preprint arXiv:2104.04670.
- 13 Schick, T., & Schütze, H. (2021). Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. arXiv preprint arXiv:2001.07676.
- 14 Gao, T., Fisch, A., & Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. arXiv preprint arXiv:2012.15723.
- 15 Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv preprint arXiv:2101.00190.

16 Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education*, 9, 1386075.

17 Zou, X., et al. (2023). Segment Everything Everywhere All at Once. In *Advances in Neural Information Processing Systems* 36.

18 Wang, W., et al. (2023). VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. In *Advances in Neural Information Processing Systems* 36.

19 Yang, L., et al. (2023). Fine-Grained Visual Prompting. In *Advances in Neural Information Processing Systems* 36.

20 Hsu, J., et al. (2023). What's Left? Concept Grounding with Logic-Enhanced Foundation Models. In *Advances in Neural Information Processing Systems* 36

## ТҮЙІН

Промпт-инжиниринг (подсказба инженериясы) ірі тілдік модельдермен (LLM) жұмыс істеудің ажырамас бөлігіне айналды, әсіресе нөлдік (zero-shot) және аз-мысалды (few-shot) оқыту сценарийлерінде, мұнда модельдер тапсырмаларды үлгілерсіз немесе ең аз санмен орындауы тиіс. Бұл мақалада математикалық пайымдау тапсырмаларындағы LLM дәлдігін арттыру тиімділігіне баса назар аудара отырып, түрлі промпт-инжиниринг әдістеріне жүйелі шолу және салыстырмалы талдау ұсынылады. Талдау үшін GSM8K, SVAMP және AQuA сияқты стандартты деректер жиынтықтары таңдалды. Әдістерге нөлдік оқыту, аз-мысалды оқыту, ойлау тізбегі (CoT), өз-өзімен келісімділік (self-consistency) және басқалар кіреді. Негізгі нәтижелер CoT-тың, әсіресе өз-өзімен келісімділікпен бірге қолданылғанда, дәлдікті едәуір арттыратынын көрсетеді; бұл, әсіресе PaLM-540B сияқты ірі модельдерде айқын: GSM8K жиынтығында дәлдік 25,1 %-дан 74,4 %-ға дейін өсті. Зерттеу нақты тапсырмаларға сай әдістерді таңдауда зерттеушілер мен практиктерге құнды ұсынымдар береді. Бұған қоса, мультимодальды жүйелердегі гибриді промпттарды қолдану перспективалары, сондай-ақ когнитивтік іздеу мен өзін-өзі бейімдеп оқыту техникасын білім беру мен өнеркәсіптік сценарийлерге біріктіру мүмкіндіктері қарастырылып, бұл қорытындылар интерпретабельдігін арттырып, ЖИ-технологияларды енгізуді жеделдетуге жол ашады.