УДК 004.912

Васильев А.Д.,
студент магистратуры, группа ФОМ-240510
Кафедра информационных систем и технологий
Уральский федеральный университет имени первого
Президента России Б.Н. Ельцина
Россия, г. Екатеринбург

FEW-SHOT ПРОМПТИНГ КАК МЕТОД АДАПТАЦИИ ЯЗЫКОВЫХ МОДЕЛЕЙ К НОВЫМ ТИПАМ ДОКУМЕНТОВ БЕЗ ДООБУЧЕНИЯ

Аннотация: *В статье рассматривается метод few-shot промптинга как способ адаптации больших языковых моделей к задаче извлечения структурированных данных из документов произвольного формата без необходимости дообучения. Проводится сравнительный анализ подходов zero-shot, few-shot и fine-tuning по критериям стоимости, скорости адаптации и качества результатов. Описывается механизм формирования промпта с примерами и его ограничения, связанные с размером контекстного окна. Предлагается интеграция few-shot промптинга с RAG-архитектурой (Retrieval-Augmented Generation) для преодоления контекстных ограничений. Приводятся результаты апробации метода на реальных корпоративных документах в рамках разработки графической утилиты для работы с OCR-сервисом.*

Ключевые слова: *few-shot промптинг, языковые модели, извлечение данных, OCR, RAG, обработка документов, LLM.*

Annotation: *The article examines few-shot prompting as a method for adapting large language models to the task of structured data extraction from documents of arbitrary format without fine-tuning. A comparative analysis of zero-shot, few-shot, and fine-tuning approaches is conducted across criteria of cost,*

adaptation speed, and output quality. The mechanism of prompt construction with examples is described, along with its limitations related to context window size. Integration of few-shot prompting with RAG architecture (Retrieval-Augmented Generation) is proposed to overcome context limitations. Results of the method's evaluation on real corporate documents within the scope of developing a graphical utility for an OCR service are presented.

Key words: *few-shot prompting, language models, data extraction, OCR, RAG, document processing, LLM.*

Введение

Автоматизация извлечения структурированных данных из неструктурированных документов является одной из приоритетных задач в современном корпоративном документообороте. Банки, страховые компании, логистические операторы ежедневно обрабатывают тысячи счетов, договоров, накладных и актов — документов, содержащих критически важную информацию, которую необходимо извлечь и передать в информационные системы организации. Ручная обработка таких объёмов является экономически нецелесообразной, а традиционные методы машинного обучения требуют значительных ресурсов на разметку и переобучение при появлении новых форматов документов.

Развитие больших языковых моделей (LLM) открыло новые возможности для решения этой задачи. Однако прямое применение LLM к выводу OCR-систем без специальной подготовки контекста демонстрирует нестабильные результаты: модель некорректно сопоставляет поля, теряет семантические связи и при большом объёме документа переполняет контекстное окно [1]. Актуальным становится вопрос о методах адаптации языковых моделей к конкретным форматам документов с минимальными затратами и без процедуры дообучения.

Цель настоящей статьи — исследовать метод few-shot промптинга как механизм адаптации LLM к задаче извлечения структурированных данных, определить его преимущества и ограничения, а также предложить архитектурное решение для его масштабирования на корпоративный документооборот.

1. Подходы к адаптации языковых моделей

В задачах извлечения информации из документов применяются три основных подхода к работе с языковыми моделями: zero-shot промптинг, few-shot промптинг и fine-tuning (дообучение). Каждый из них характеризуется различным соотношением затрат на подготовку и качества результата.

Zero-shot промптинг предполагает подачу задания модели без каких-либо примеров ожидаемого результата. Модель опирается исключительно на знания, накопленные в процессе предобучения. Для хорошо стандартизированных форматов (например, общеизвестных типов счетов-фактур) этот подход может давать приемлемые результаты, однако при работе с уникальными корпоративными шаблонами качество извлечения резко снижается из-за отсутствия контекста о структуре конкретного документа [2].

Fine-tuning — процедура дополнительного обучения предобученной модели на специализированном датасете — обеспечивает высокое качество извлечения данных для фиксированного набора форматов документов. Вместе с тем данный подход имеет существенные ограничения: для каждого нового типа документа требуется формирование размеченного датасета объёмом от нескольких сотен до тысяч примеров, что занимает значительное время и требует высокой квалификации разметчиков. Кроме того, адаптация модели к изменившемуся шаблону существующего документа требует повторного обучения [3].

Few-shot промптинг занимает промежуточное положение: в промпт, подаваемый модели, включается небольшое количество пар «вход —

ожидаемый выход», демонстрирующих правильный формат извлечения. Такой подход позволяет адаптировать модель к новому типу документа. Ниже приведено сравнение трёх подходов по ключевым критериям.

Таблица 1.

Сравнительный анализ подходов к адаптации языковых моделей

Критерий	Zero-shot	Few-shot	Fine-tuning
Необходимость обучающих данных	Не требуется	Несколько примеров (3–10)	Сотни/тысячи примеров
Стоимость адаптации	Минимальная	Низкая	Высокая
Качество извлечения	Нестабильное	Высокое при правильных примерах	Высокое при достаточных данных
Скорость адаптации к новому типу документа	Мгновенная	Минуты	Дни–недели
Ограничения	Галлюцинации, ошибки структуры	Размер контекстного окна	Зависимость от объёма разметки

Как видно из таблицы 1, few-shot промптинг демонстрирует оптимальный баланс между затратами на адаптацию и качеством результатов, что делает его наиболее привлекательным для корпоративного применения в условиях постоянно меняющегося пула форматов входящих документов.

2. Механизм few-shot промптинга при извлечении данных из документов

Few-shot промптинг основан на принципе обучения на примерах в контексте (in-context learning). Промпт, подаваемый языковой модели, имеет следующую структуру: системная инструкция, описывающая задачу и формат вывода; один или несколько блоков «пример входных данных — пример правильного вывода»; целевые входные данные, для которых требуется получить результат.

В контексте извлечения данных из документов входными данными является текстовый вывод OCR-системы — неструктурированный поток слов с координатами. Ожидаемым выводом выступает JSON-объект, содержащий именованные поля документа и их значения. Каждый пример в промпте показывает модели, каким образом необходимо сопоставлять текстовые фрагменты конкретного типа документа с полями целевой схемы [4].

Критическим параметром является качество примеров. Исследования показывают, что релевантность примера целевому документу (схожесть структуры, типа полей, способа оформления) влияет на точность извлечения в большей степени, чем количество примеров [2]. Это обстоятельство обусловливает необходимость механизма динамического выбора наиболее подходящего примера из накопленной базы шаблонов для каждого нового входящего документа.

Ключевым ограничением подхода является размер контекстного окна языковой модели. При работе с объёмными документами (многостраничные договоры, детализированные счета) суммарный объём системного промпта, примеров и целевого текста может превышать допустимый лимит токенов, что приводит к усечению входных данных и потере информации. Данная проблема требует архитектурного решения на уровне системы управления шаблонами.

3. RAG-архитектура как решение проблемы масштабирования

Для преодоления контекстных ограничений и обеспечения масштабируемости системы на большой пул форматов документов предлагается интеграция few-shot промптинга с архитектурой Retrieval-Augmented Generation (RAG). В этой модели база шаблонов хранится не в промпте целиком, а во внешнем векторном хранилище. Перед формированием запроса к языковой модели система выполняет поиск наиболее релевантного шаблона для поступившего документа.

Архитектура включает следующие компоненты: модуль предобработки OCR-вывода, преобразующий XML-схему с координатами слов в плоский текст с восстановленной структурой; векторный индекс шаблонов, в котором каждый шаблон представлен вектором эмбеддинга, вычисленным по совокупности входного текста и целевых полей; модуль семантического поиска, выбирающий из индекса k ближайших шаблонов к текущему документу; модуль сборки промпта, формирующий итоговый запрос из системной инструкции, отобранных примеров и целевого текста [5].

Принципиальным преимуществом данного подхода является то, что в промпт попадает лишь один-два наиболее релевантных шаблона вместо всей базы, что кардинально сокращает потребление контекстного окна. При этом общее количество поддерживаемых типов документов ограничено только ёмкостью векторного хранилища, а не размером контекста модели. Накопление новых шаблонов не требует переобучения — достаточно добавить размеченный пример в индекс.

4. Апробация метода

Описанный метод был реализован в рамках разработки графической утилиты для работы с OCR-сервисом «Протон» (компания ООО «Нексус»). Утилита функционирует в двух режимах: режим обучения, в котором пользователь интерактивно размечает примеры путём выделения фрагментов документа и назначения им имён полей; режим распознавания, в котором новый документ обрабатывается автоматически с использованием RAG-поиска и few-shot промптинга.

Апробация проводилась на выборке из 120 документов четырёх типов: счета-фактуры, товарные накладные, акты выполненных работ и договоры поставки. Для каждого типа был сформирован пул из 5 размеченных примеров. В качестве языковой модели использовалась локально развёрнутая модель с контекстным окном 8 000 токенов.

Точность извлечения ключевых полей (наименование контрагента, сумма, дата, номер документа) составила 94,2% для стандартных форматов документов и 87,6% для документов с нестандартным расположением полей. Для сравнения, тот же набор документов, обработанный без примеров (zero-shot), показал точность 71,3%, что подтверждает существенное преимущество few-shot подхода.

В ходе апробации с участием фокус-группы из 8 специалистов по документообороту было установлено, что среднее время создания нового шаблона для ранее неизвестного типа документа составило 4,5 минуты, что кратно превосходит по скорости процедуру дообучения, требующую нескольких дней.

Заключение

Проведённое исследование демонстрирует, что few-shot промптинг является эффективным методом адаптации языковых моделей к задаче извлечения структурированных данных из документов произвольного формата. Метод обеспечивает высокое качество извлечения при минимальных затратах на подготовку — достаточно разметить 3–10 примеров для нового типа документа.

Интеграция few-shot промптинга с RAG-архитектурой позволяет преодолеть ключевое ограничение метода — размер контекстного окна — и обеспечивает масштабируемость системы на неограниченное количество форматов документов без переобучения языковой модели. Апробация метода на реальных корпоративных документах подтвердила его практическую применимость: точность извлечения полей составила до 94,2% при времени создания нового шаблона менее 5 минут.

Перспективным направлением дальнейших исследований является разработка методов автоматической валидации результатов извлечения и механизмов активного обучения, при которых система самостоятельно

выявляет низкоуверенные результаты и запрашивает разметку дополнительных примеров у оператора.

Использованные источники:

1. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 9459–9474.

2. Min S., Lyu X., Holtzman A., et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — 2022. — P. 11048–11064.

3. Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models // International Conference on Learning Representations. — 2022. — URL: https://arxiv.org/abs/2106.09685 (дата обращения: 15.04.2025).

4. Brown T., Mann B., Ryder N., et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 1877–1901.

5. Gao Y., Xiong Y., Gao X., et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv. — 2024. — URL: https://arxiv.org/abs/2312.10997 (дата обращения: 15.04.2025).