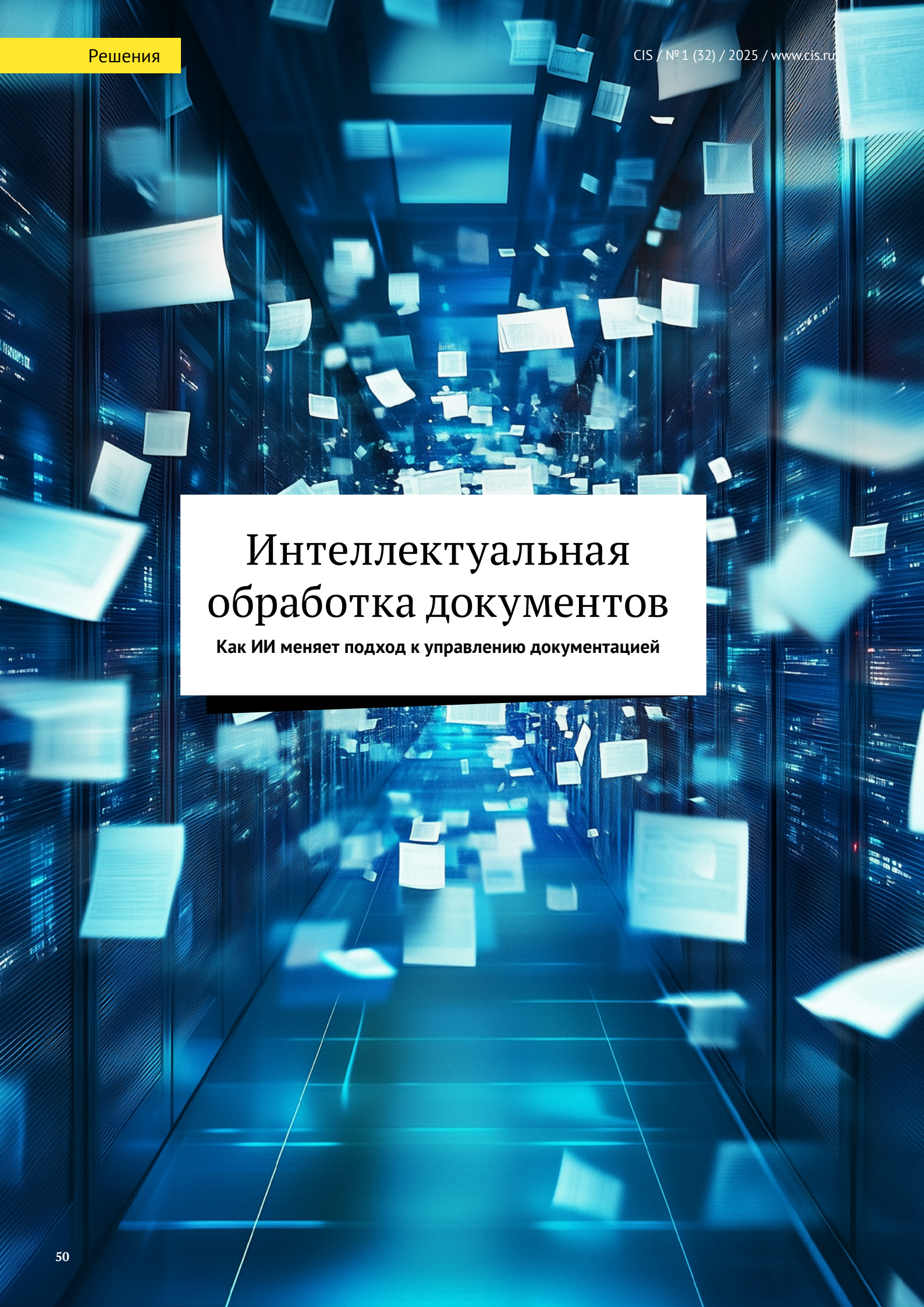


Интеллектуальная обработка документов

Как ИИ меняет подход к управлению документацией

О роли ИИ в управлении документами рассказывает директор департамента бизнес-решений компании «Инфолоджистикс» Павел Гаврилов.

Значимость технологий, основанных на нейронных сетях, специализирующихся на обработке естественного языка, или больших лингвистических моделей (Large language model, LLM), не требует дополнительного упоминания. Присутствующий вокруг этой темы информационный бум последних лет уже сделал эту работу.

Технологии генеративного ИИ (GenAI), согласно исследованию Gartner в 2024 году, находятся в завершающей стадии этапа завышенных ожиданий, где новая технология ещё воспринимается как «волшебная палочка», решающая все проблемы здесь и сейчас (рис. 1).

Компания «Инфолоджистикс» производит большую работу по обоснованным применениям ИИ, уделяя особое внимание сфере аутсорсинга решений и сервисов, связанных с управлением документами и информацией.

Предлагаем читателю квант осмысления исследовательской работы в применении GenAI к задачам работы с документами. Также примем на себя роль прогнозиста, чтобы сформулировать следствия, полезные другим.

«I» – часть от «AI» – когда «всё сам»

В предыдущие годы взаимодействие с системами СЭД и ECM происходило в парадигме, «что вложил, то и получил». Зачастую пользователи вносили данные о добавляемых объектах в системы вручную. Настройки систем были жёстко фиксированными либо допускали небольшие отклонения от заданных правил, прописанных «в коде» или конструкторах маршрутов.

Развитие функциональности конструкторов давало сделать больше, что порой приводило к очень сложным реализациям логики движения данных. Ниже пример бизнес-логики для всего лишь согласования служебной записки в одной из компаний.

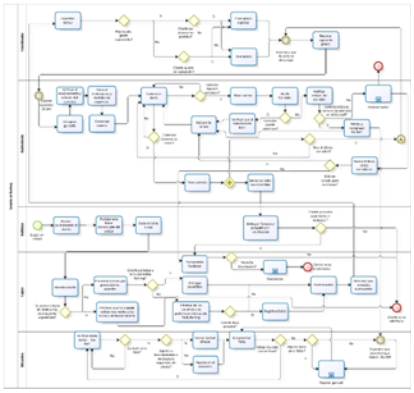
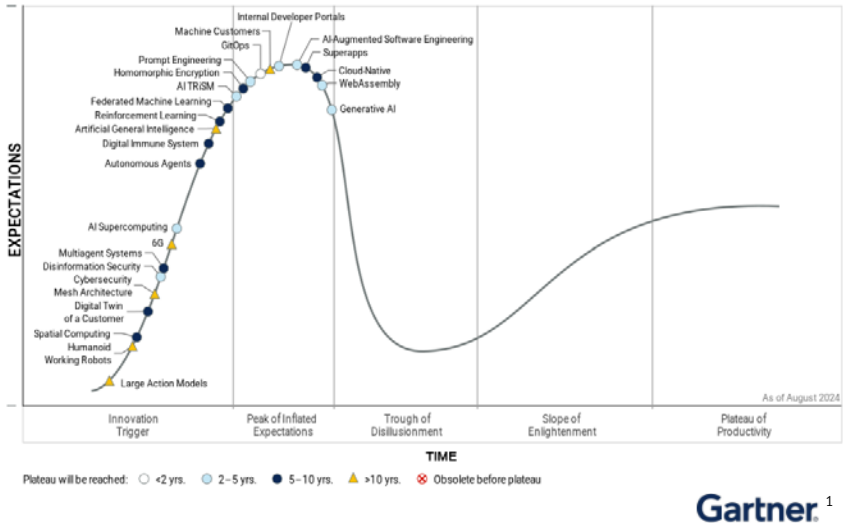


Рисунок 1



Со временем в цепочках сменяются ответственные, изменяются реальные бизнес-процессы, происходят разрывы преемственности команд. Поддержание логик усложняется и требует ресурсов, которые компании не всегда могут выделить.

Есть и другая проблематика, связанная с документооборотом, – попытка решить задачу ввода данных в системы и характеризовать создаваемые в них объекты набором атрибутивной информации (номер и дата документа, наличие нужных подписей и т. д.).

Такие решения реализовывались через дорогостоящие системы OCR, где создавались порой сотни предварительно заданных форм документов, без чего процесс распознавания не работал корректно, формы документов должны быть известны и настроены заранее. Сами формальные правила обработки документов также меняются со временем, что приводит к потребности перевнедрения.

Даже хорошие внедрения требовали людей «в петле» процесса для верификации распознанных данных документов плохого качества и не подходящих под распознавание. Как следствие, не достигалась цель полной автоматизации, что нивелировало экономический эффект проекта. Всё это характерно для внедрений тяжеловесных решений по распознаванию, сделанных в районе 2010-2020 годов.

Суммируем проблемы, связанные с устоявшимися до последнего времени решениями.

- Ручной ввод данных.
- Шаблонное распознавание и необходимость постоянного пополнения альбома форм.
- Ручная верификация распознанных данных.
- Сложность формулировки и сопровождения бизнес-правил.
- Малая осведомлённость о новых технологиях и подходах.

Сложившаяся ситуация привела нас к попытке решить эти проблемы через применение ИИ, информация о гипотетических возможностях которого есть повсюду.

1. https://emt.gartnerweb.com/ngw/globalassets/en/newsroom/images/graphs/august_2024_ethc.png

«AI» – ИИ по 2022 год

Мы могли бы начать отсчёт возрождения нейронных сетей в форме GenAI с идеи, описанной в 2009 году в статье Google «The Unreasonable Effectiveness of Data»[1] (Google, 2009). Исследователи отмечали, что обучение сетей на больших объёмах неразмеченных данных – эффективнее, чем на меньших, но предварительно структурированных.

Существовавшие ранее на рынке системы, созданные на основе правил и знаний, собираемых и пополняемых вручную большим количеством экспертов, стали быстро терять актуальность. Появились продукты, которые регулярно пополняют свои базы данных и автоматически формируют правила соотнесения данных внутри себя. Такие принципы работы распространились и на документационные системы, что привело к развитию некоторых особенностей.

- Строгие шаблоны уступили место гибким, возможность находить нужные поля возросла.
- Распознавание охватило не только текст, но и фото.
- Нахождение особых паттернов в медицинских данных (медицинские снимки).
- Возросла взаимоувязанность систем распознавания и систем-источников и -приёмников ввиду общего развития сред разработки и API.
- Меньше человеческого труда для ввода данных. Появились доступные решения для типичных задач в учётных/WMS системах.

1. https://research.google.com/pubs/archive/35179.pdf

AI + (AI×AI×AI)= AI^10

Что нас ждёт за горизонтом?

Исследователи из Epoch[2], занимающиеся анализом и прогнозированием развития передовых систем ИИ, в работе «Will we run out of data?» прогнозируют, что вскоре производимой человечеством информации уже не будет достаточно для дальнейшего обучения моделей.

Artificial general intelligence (AGI) – тот самый настоящий искусственный интеллект, равный или превосходящий человеческий в части обработки и синтеза новой информации. Предполагается, что после исчерпания данных для обучения моделей, их будут создавать другие генеративные модели, что и откроет путь к созданию AGI (рис. 2).

Какие выводы можно сделать прямо сейчас?

1. **Накапливайте все возможные данные о деятельности компании:**
   - документы;
   - внешнюю и внутреннюю переписку;
   - записи разговоров;
   - логи изменений в учётных системах;
   - архивы операционных данных;
   - документационные архивы.

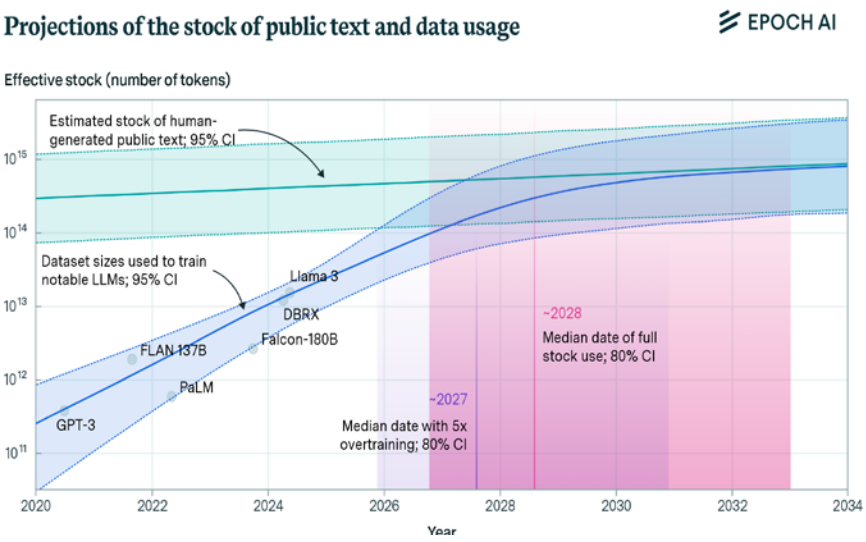
   Может наступить момент, когда ранее непонятно для чего нужные данные позволят усилить внедряемые вами решения, основанные на AGI. Это откроет путь к быстрому прогнозированию, анализу хозяйственной деятельности и оценке различных вариантов развития вашей компании.
2. **Неструктурированных данных практически не существует.** Зачастую они так или иначе структурированы: есть иерархии заголовков, глав, перекрёстных ссылок, форматов и графиков, нумерации страниц, метки времени и другие метаданные, которые в будущем могут быть использованы ИИ для анализа.
3. **Откройте заново ценность архивов.** Накопленные документационные архивы первичных документов, техническая документация, дела по финансовым продуктам, данные геологоразведки и так далее – действительно, за прошедшие годы накоплено огромное количество данных, которые можно пустить в дело только одним способом: оцифровать их.

2. https://epoch.ai

Экономически целесообразно делать это сейчас, чем позже пытаться сделать такую работу одномоментно в будущем. Зачастую при выполнении таких проектов уже сейчас графики работы строятся на годы вперёд ввиду общей трудоёмкости задачи.

Оцифровка и структурирование архива – не только решение текущих задач, но и залог для будущих возможностей формировать ценные выводы о деятельности компании для создания преимущественного позиционирования на рынке.

Рисунок 2



INFOLOGISTICS
BUSINESS PROCESS AUTOMATION

«Инфолоджистикс»
infologistics.ru