От поиска к пониманию: эволюция RAG — как превратить корпоративные знания в интеллектуального собеседника

Жарков Олег

Ростелеком

ПОЧЕМУ LLM НЕ МОЖЕТ БЫТЬ БЕЗ RAG?

01

Галлюцинации = ущерб бизнесу

- Юридические риски
- Финансовые риски
- Репутационные риски

02

Устаревание знаний

03

Модели недоступны внутренние данные компании

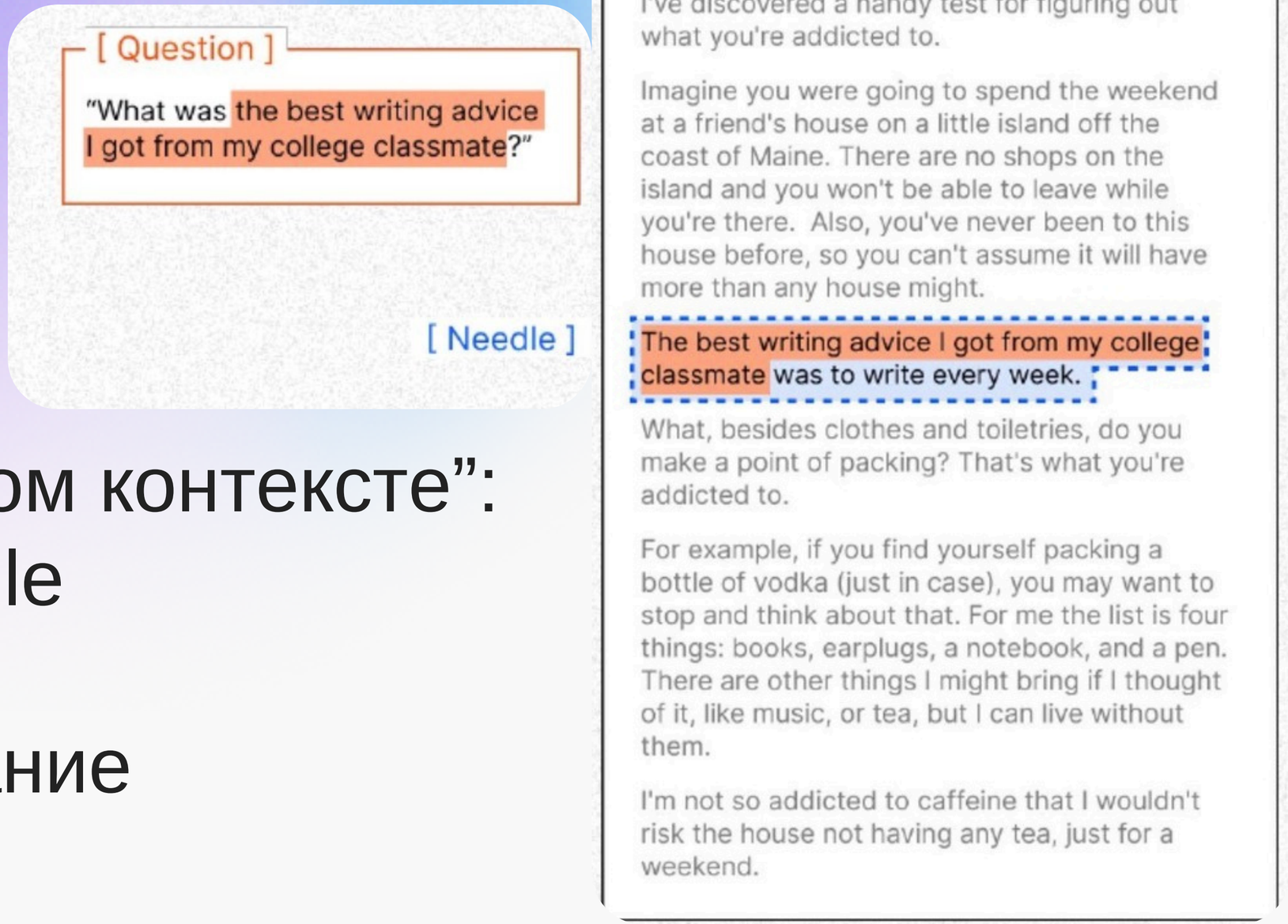
04

Отсутствие прозрачности и аудита

LONG CONTEXT WINDOW VS RAG

> Зачем сейчас нужен RAG, если у моделей уже контекст на миллион токенов? Давайте просто положим туда все документы.

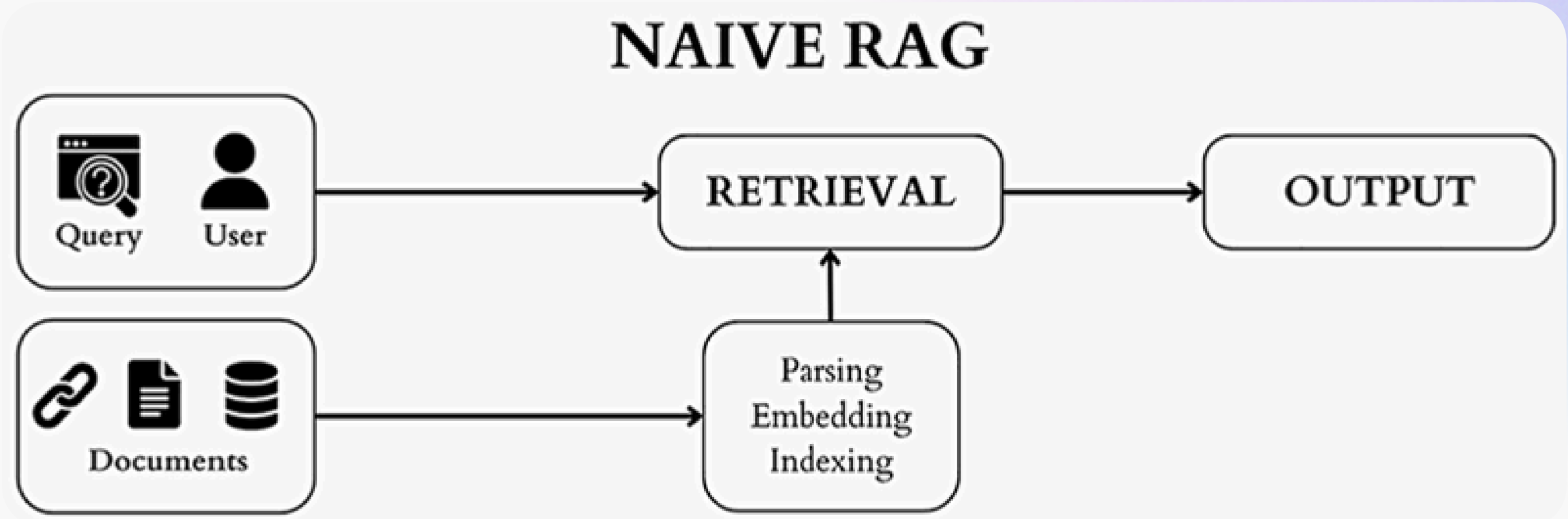
LONG CONTEXT WINDOW VS RAG

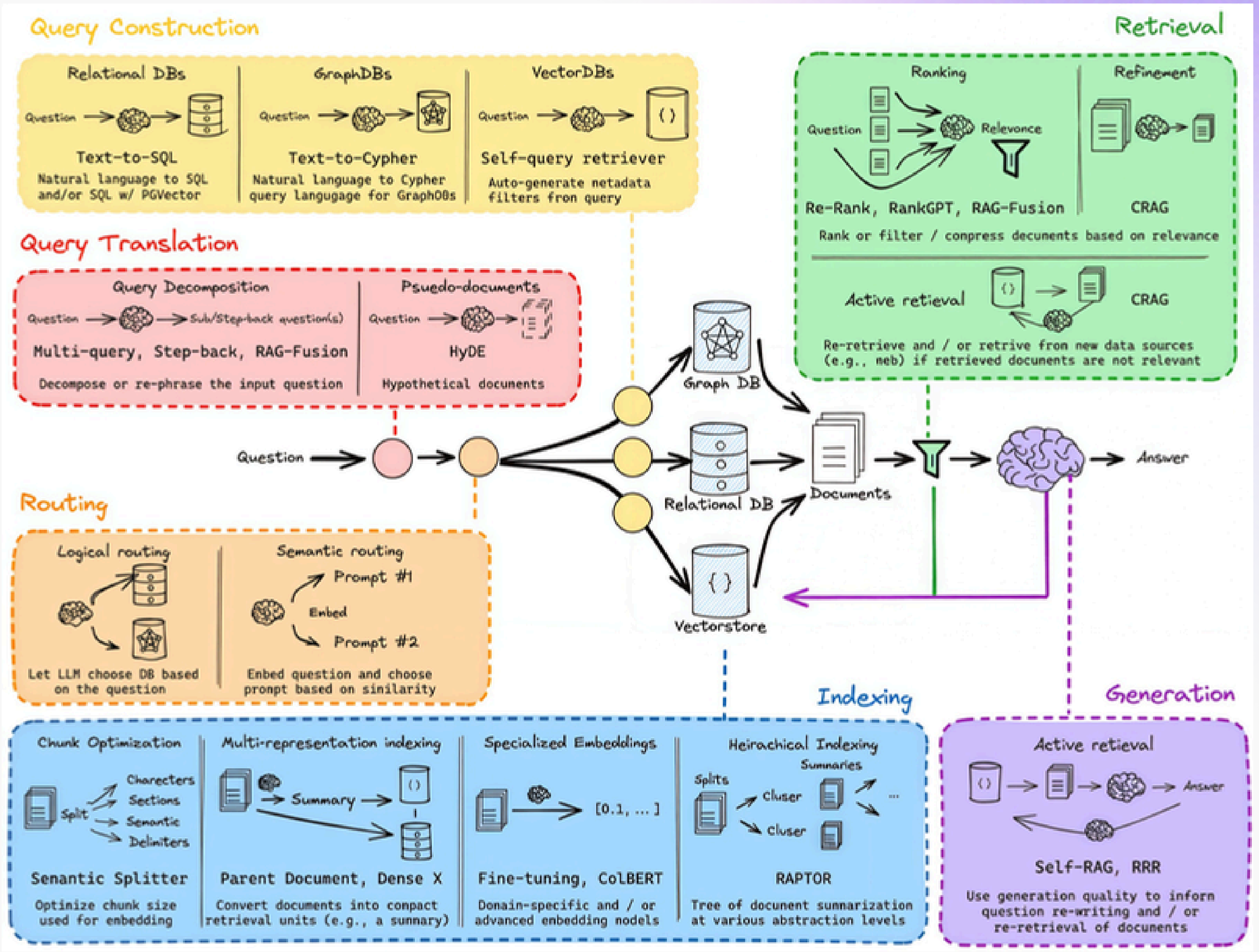


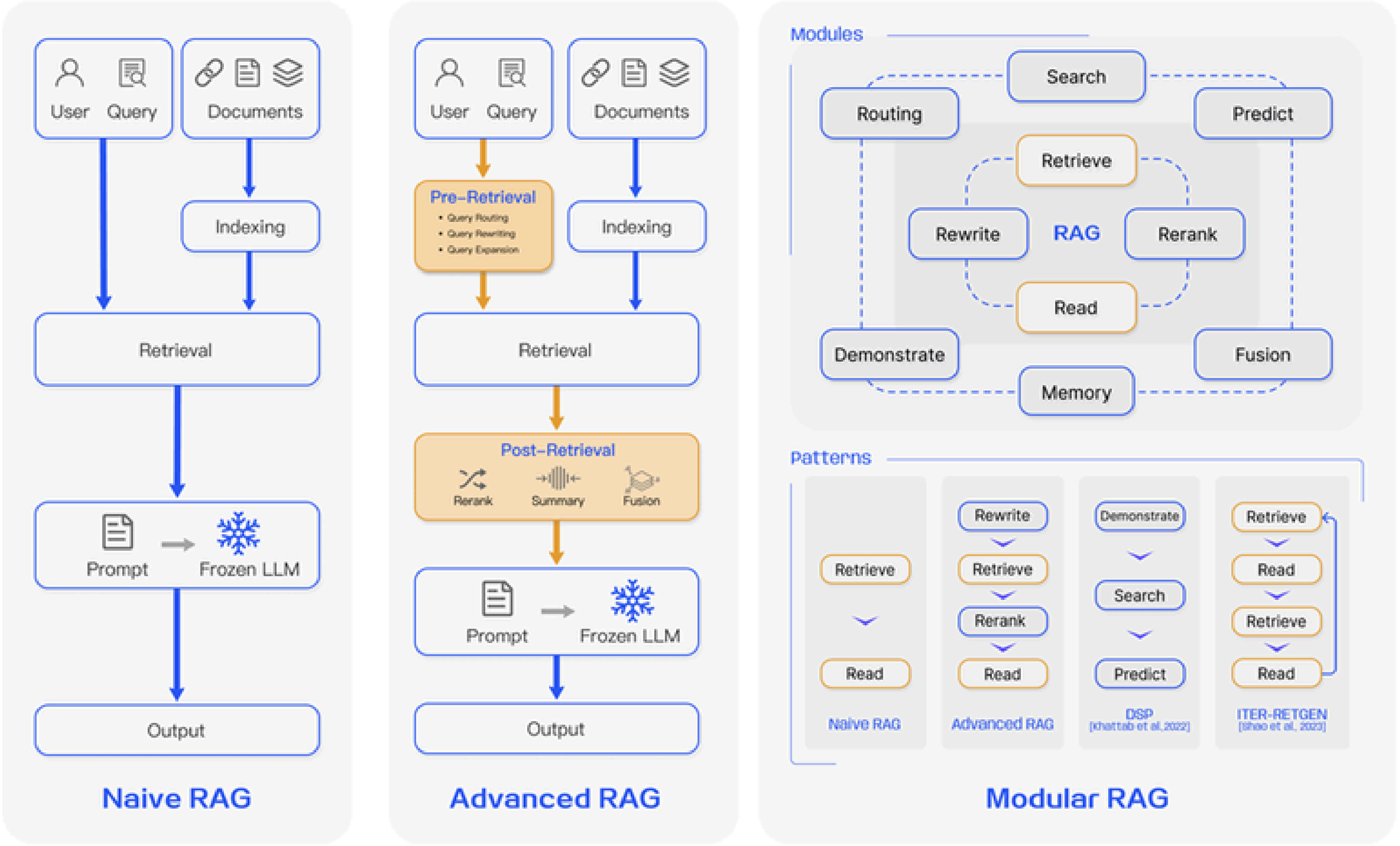
Миф о “бесконечном контексте”:

- Lost in the middle
- Цена и latency
- Масштабирование

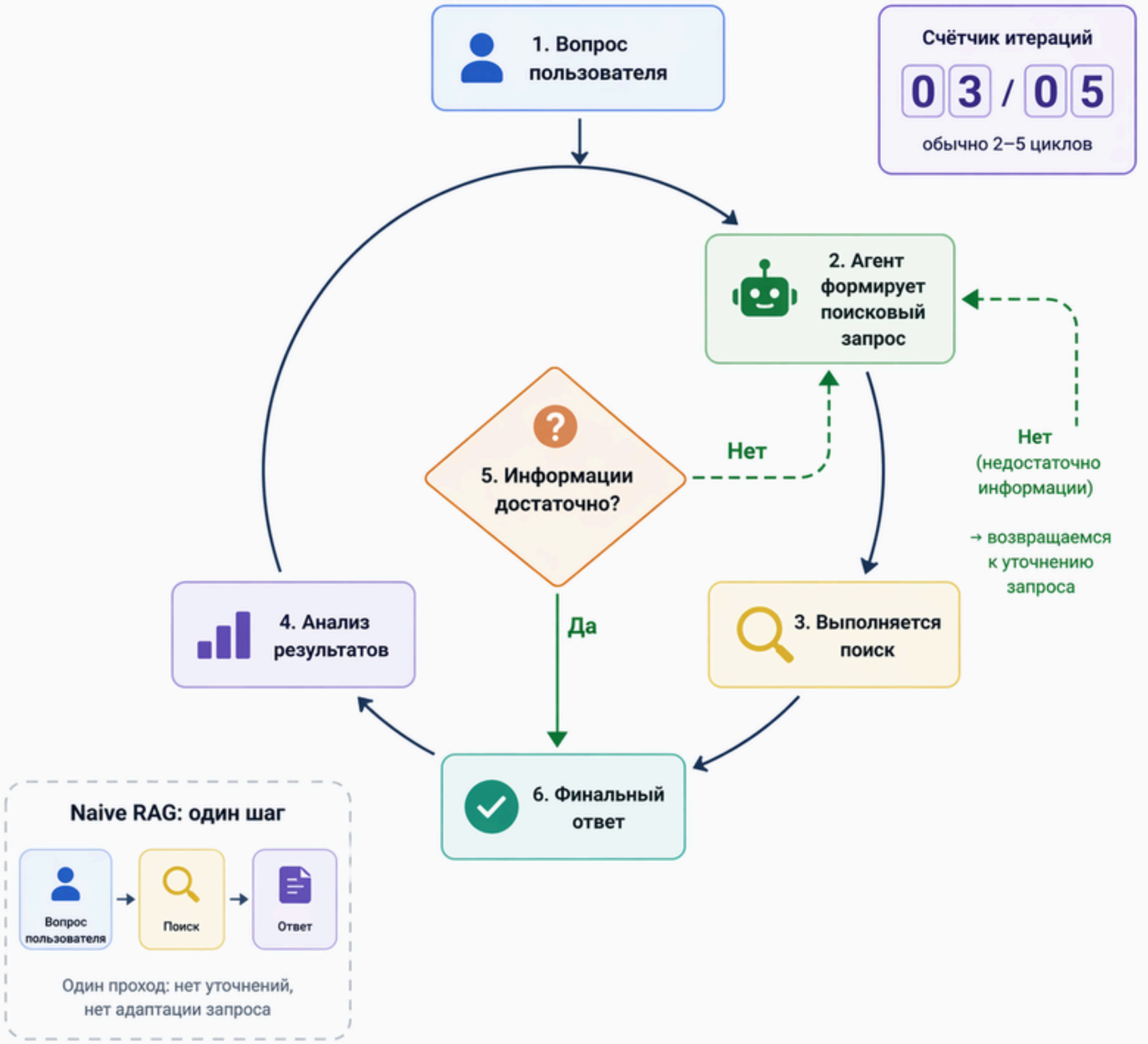
ЭВОЛЮЦИЯ АРХИТЕКТУРЫ RAG



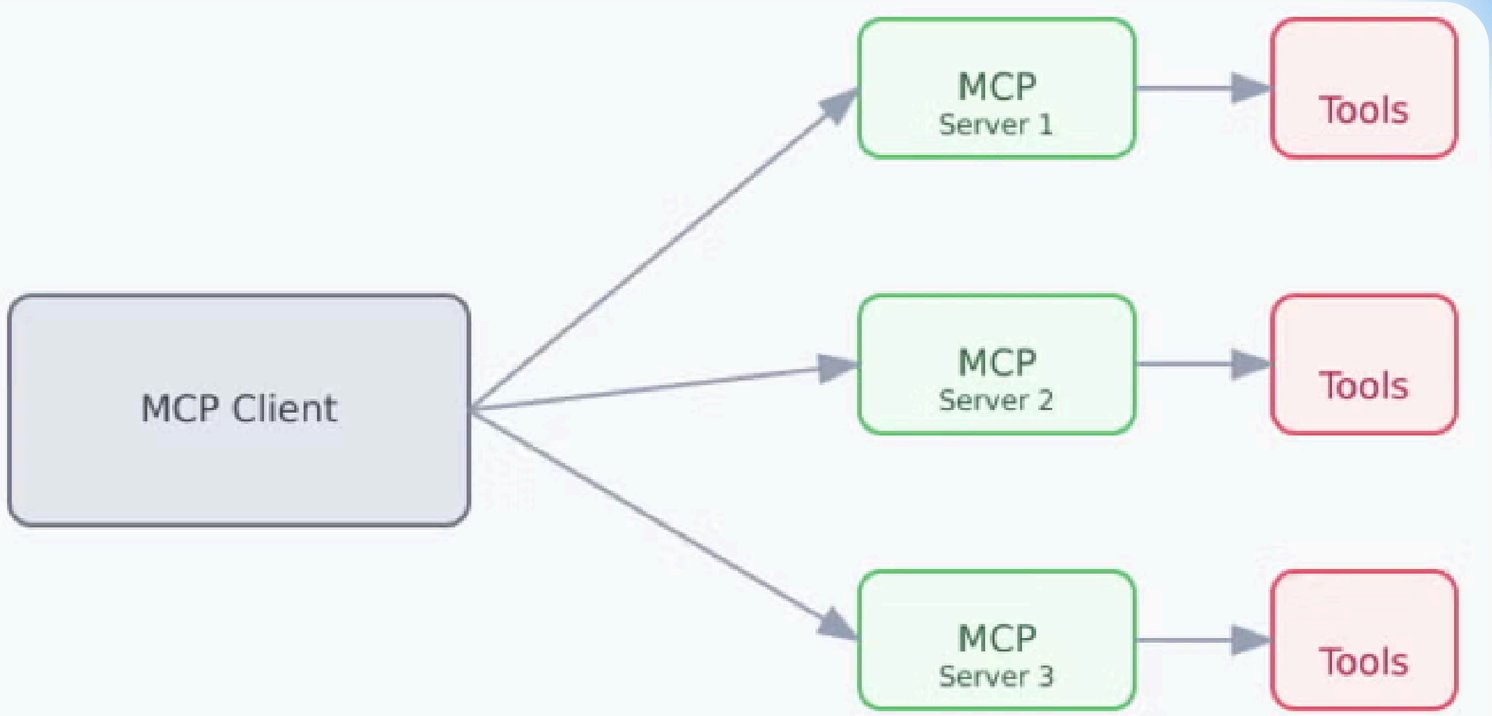

Query Construction
Relational DBs
Question
Text-to-SQL
Natural language to SQL and/or SQL w/ PGVector
GraphDBs
Question
Text-to-Cypher
Natural language to Cypher query langugage for GraphDBs
VectorDBs
Question
Self-query retriever
Auto-generate metadata filters from query
Retrieval
Ranking
Question
Relevance
Re-Rank, RankGPT, RAG-Fusion
Refinement
CRAG
Rank or filter / conpress decuments based on relevance
Active retieval
CRAG
Re-retrieve and / or retrive from new data sources (e.g., web) if retrieved documents are not relevant
Query Translation
Query Decomposition
Question
Sub/Step-back question(s)
Multi-query, Step-back, RAG-Fusion
Decompose or re-phrase the input question
Psuedo-documents
Question
HyDE
Hypothetical documents
Graph DB
Relational DB
Vectorstore
Question
Documents
Answer
Routing
Logical routing
Let LLM choose DB based on the question
Semantic routing
Prompt #1
Embed
Prompt #2
Enbed question and choose prompt based on similarity
Indexing
Chunk Optimization
Split
Characters
Sections
Semantic
Delimiters
Senantic Splitter
Optimize chunk size used for embedding
Multi-representation indexing
Summary
Parent Document, Dense X
Convert documents into conpact retrieval units (e.g., a summary)
Specialized Embeddings
[0.1, ...]
Fine-tuning, ColBERT
Donain-specific and / or advanced enbedding nodels
Heirachical Indexing
Splits
Summaries
cluser
cluser
RAPTOR
Tree of document summarization at various abstraction levels
Generation
Active retieval
Answer
Self-RAG, RRR
Use generation quality to inform question re-writing and / or re-retrieval of documents


User
Query
Documents
Indexing
Retrieval
Prompt
Frozen LLM
Output
Naive RAG
User
Query
Documents
Pre-Retrieval
Query Routing
Query Rewriting
Query Expansion
Indexing
Retrieval
Post-Retrieval
Rerank
Summary
Fusion
Prompt
Frozen LLM
Output
Advanced RAG
Modules
Search
Routing
Predict
Retrieve
Rewrite
RAG
Rerank
Read
Demonstrate
Fusion
Memory
Patterns
Retrieve
Read
Naive RAG
Rewrite
Retrieve
Rerank
Read
Advanced RAG
Demonstrate
Search
Predict
DSP
[Khattab et al.2022]
Retrieve
Read
Retrieve
Read
ITER-RETGEN
[Shao et al. 2023]
Modular RAG

Agentic RAG — итеративный цикл



MCP (Model Context Protocol) — как новый стандарт интеграции RAG в агентные системы



На практике — три эффекта:

- Ускорение интеграций
- Стандартизация retrieval-слоя
- Единая политика доступа и аудита

БЕЗОПАСНОСТЬ ДОВЕРИЕ И КОНТРОЛЬ RAG

Когда данные не могут покидать контур:

- персональные данные, коммерческая тайна, NDA;
- отрасли с жёсткой регуляторикой (госсектор, медицина, финансы, телеком);
- внутренние инциденты, уязвимости, безопасность.

GUARDRAILS: ЗАЩИТА ОТ PROMPT-ИНЪЕКЦИЙ И НЕГАТИВНЫХ ТЕМ

01 Direct injection

02 Indirect injection

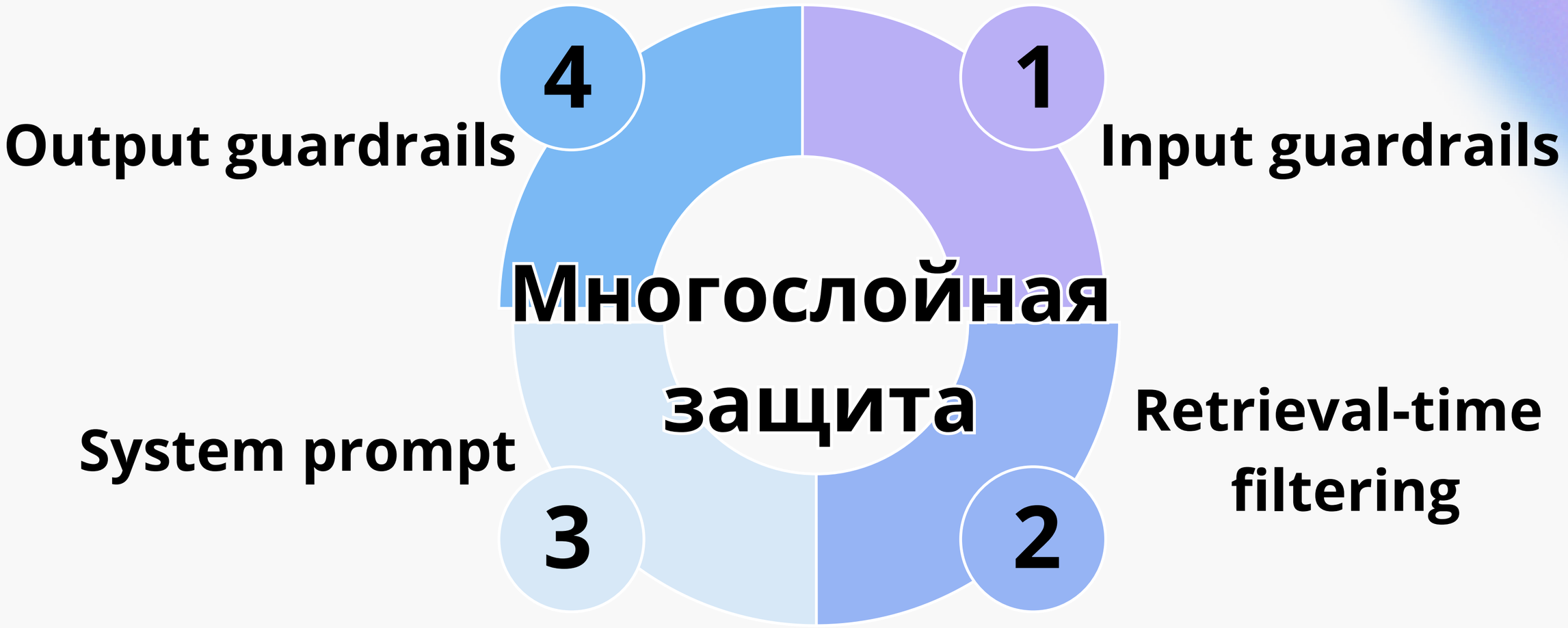
03 Jailbreak

04 Prompt leaking

05 Multimodal injection

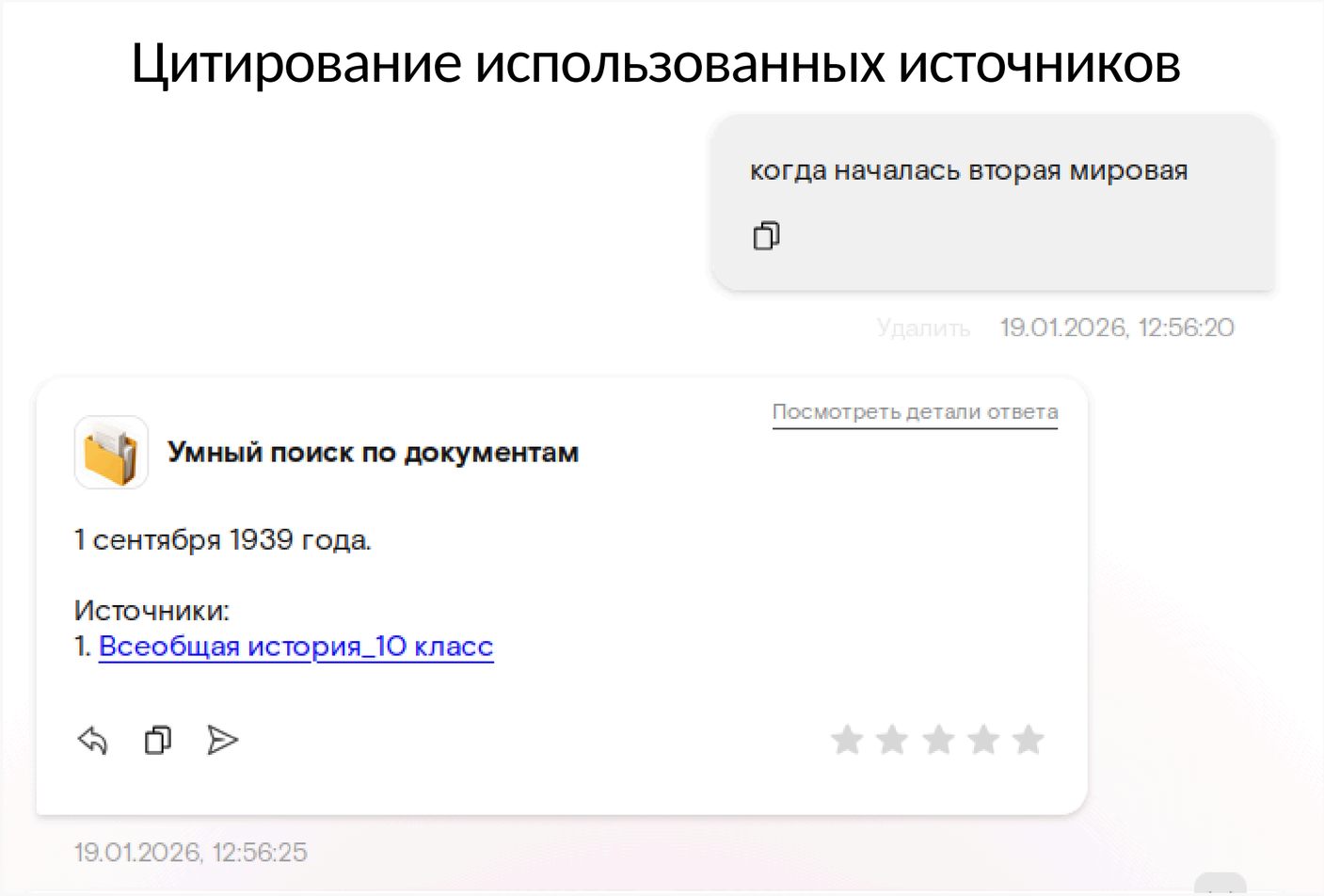
06 Cross-Tenant Leakage

GUARDRAILS: ЗАЩИТА ОТ PROMPT-ИНЪЕКЦИЙ И НЕГАТИВНЫХ ТЕМ



OBSERVABILITY ДЛЯ RAG

Цитирование использованных источников



МЕТРИКИ ОЦЕНКИ RAG

↓

Recall@K

Precision@K

MRR (Mean Reciprocal Rank)

NDCG@K / MAP

↓

Answer Relevancy

Faithfulness / Groundedness

Answer Correctness

Context Recall

Context Precision

УМНЫЙ ПОИСК ПО ДОКУМЕНТАМ

Показатели I квартала 2026г.:

50 000

пользователей платформы

10 000

активных пользователей

20 000

поисковых запросов в месяц в сервисе Умный поиск по документам

40%

ожидаемый прирост в 2026г.

ПЛАНЫ РАЗВИТИЯ

Спасибо за внимание!

Ростелеком