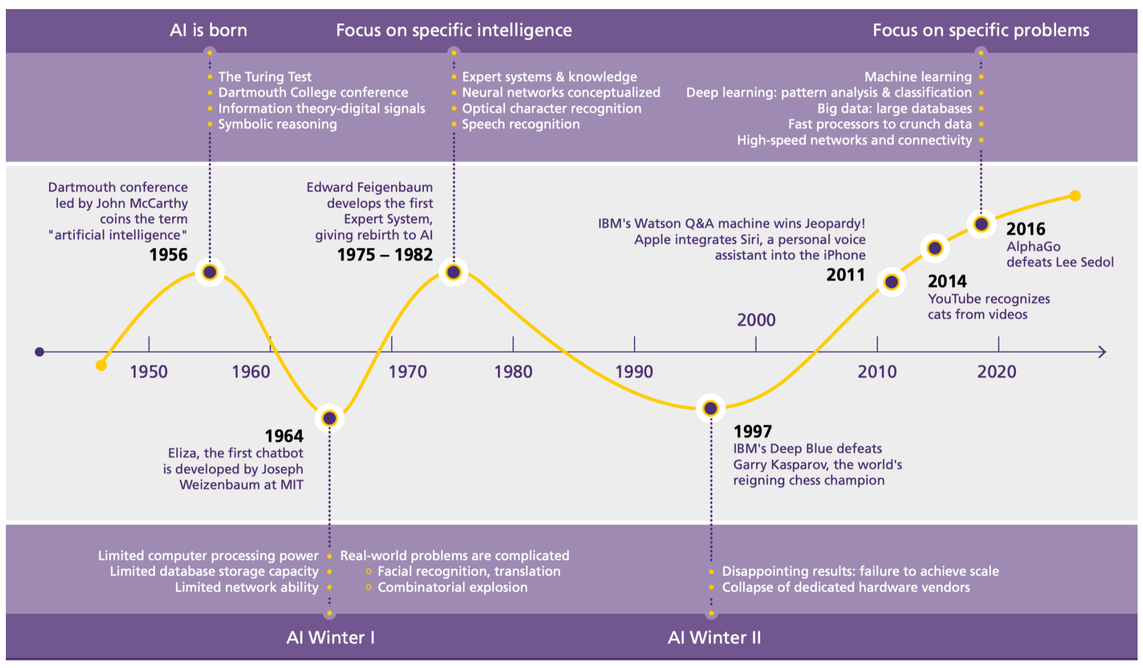
Retrieval Augmented Generation
Does the combination of pre-retrieval optimizations create added value?

AI Timeline – Optimism vs. Pessimism



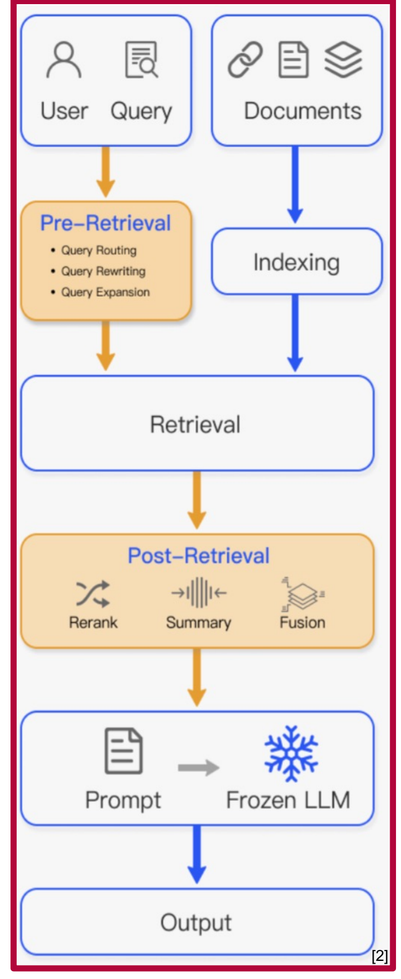
[1]

- **Eliza:** Simulated conversation by using a pattern-matching and substitution methodology to give predefined responses to users' inputs, mimicking a Rogerian psychotherapist.
- **Combinatorial explosion:** The exponential growth of possible solutions in a problem space, making it computationally infeasible to solve as the number of variables increases.
- **IBM Deep Blue:** The pioneering chess computer made history in 1997 by defeating world champion Garry Kasparov
- **Machine learning:** A branch of artificial intelligence where algorithms learn from and make predictions or decisions based on data
- **Generative AI:** Involves algorithms that can create new content, such as text, images or music by learning patterns from existing data and generating original outputs that mimic the learned data.

Making AI Reliable

Large Language Models (LLMs) demonstrate significant capabilities but face challenges such as hallucination, outdated knowledge, and non- transparent, untraceable reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the models, particularly for knowledge intensive tasks, and allows for continuous knowledge updates and integration of domain specific information [2]

Advanced RAG



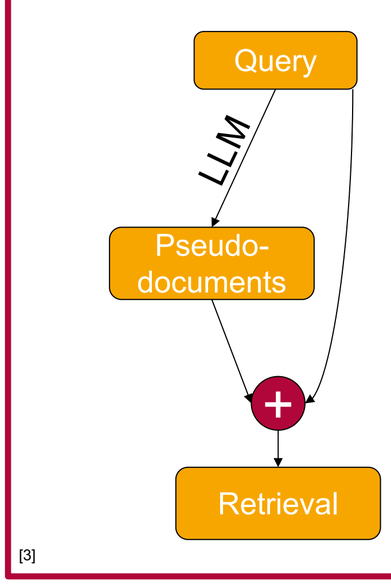
[2]

Pre-Retrieval optimization

Aligning Queries and Documents

The Users original query may suffer from imprecise phrasing and lack of semantic information. Therefore it is crucial to align the semantic space of the users query with those of the documents [2]

query2doc



[3]

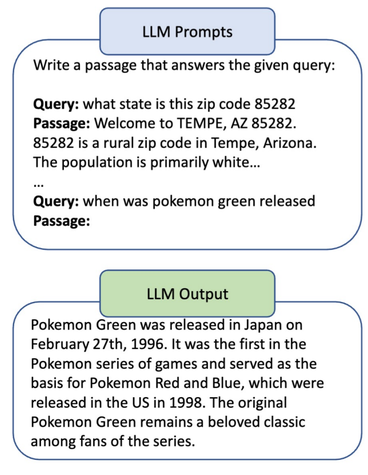
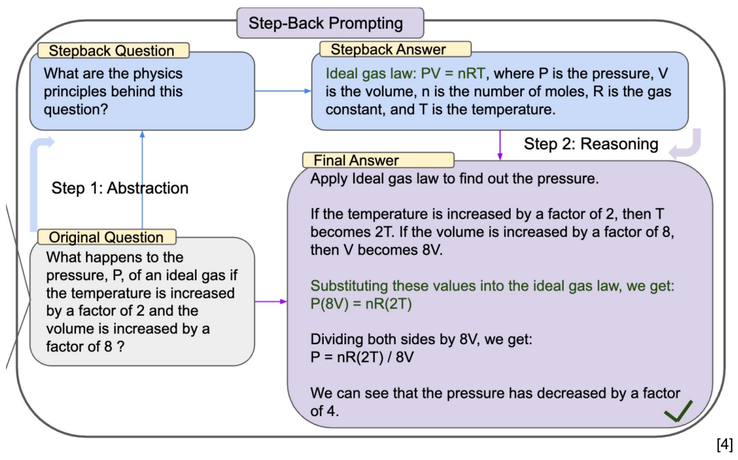


Figure 1: Illustration of *query2doc* few-shot prompting. We omit some in-context examples for space reasons.

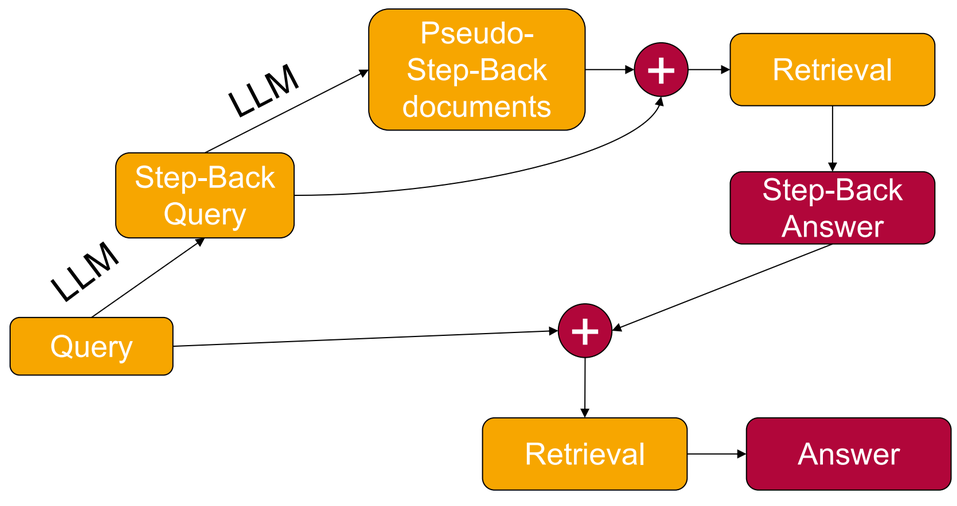
Step-Back Prompting

A simple prompting technique that enables LLMs to do abstractions to derive high-level concepts and first principles from instances containing specific details



[4]

Proposal: Step-Back query2doc



[5]

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Finn Freiheit
Ma. IT Systems Engineering
Hasso Plattner Institute
finn.freiheit@hpi.uni-potsdam.de

[1] Lavenda D. & Marsden P.
[2] Retrieval-Augmented Generation for Large Language Models: A Survey
[3] Query2doc: Query Expansion with Large Language Models
[4] TAKE A STEP BACK: EVOKING REASONING VIA AB- STRACTION IN LARGE LANGUAGE MODELS
[5] Doc2Query--: When Less is More