Fresh LLMs

① Fresh QA ⇒ A Novel Dynamic QA Benchmark encompassing a Diverse range of questions and answer types.

Including 1nQA types ⇒ fast-changing world knowledge

Pna ⇒ false premises that need to be debunked.

② Fresh Prompts ⇒ single few shot prompting method that substantially boosts the performance of an LLM on fresh QA by incorporating relevant and up to date information from search engine to the prompt.

Hallucination ⇒ Partially attributed by the performance of outdated knowledge

To mitigate this #Human feedback, or #knowledge Enhanced tasks but not easily scalable for real time knowledge updates.
(Eg: Stock Price of a company)

In Context learning = alternative in which real-time knowledge can be injected into LLM's. Prompt for Conditional Generation.

Fresh QA (600 Questions)

DATA COLLECTION

① Quality Control: Data Cleaning and Quality Assessments. Manual Review

① Well formed Questions

② Removal of Duplicates and Invalid d

③ Verification of answers and supporting evidence URL's.

New Manually Collect supplementary valid answers

② Data Size and Split:

③ Fresh QA Requires regular updates:

Updating Newer versions of data in the dataset

EVALUATION

Evaluation in 2 Modes

Relaxed

evaluate the correctness of the primary answer

Strict

additionally examine whether all the facts in the answer are accurate (No Hallucination)

① Evaluation Protocol

② Inter-rater Agreement and automatic evaluation.

{ 99% of Relaxed
96% Strict }

Additionally, FRESH EVAL simple automatic metric that uses few shot learning to

teach an LLM to judge in-context Model Responses.
acheiving agreement 96.5% of Relaxed 96% of strict for Human Responses.

FRESH PROMPT:

leverages a text prompt to

① to introduce contextually relevant and up to date information. from a Search Engine to pre-trained LLM

② Teach the Model to reason over retrieved evidences.

① Fresh Prompt Significantly improves Fresh QA Accuracy.

② FRESH PROMPT outperforms other search augmented methods by a Large Margin.

③ Premise check boosts accuracy on false premise questions but can hurt accuracy on those with valid premises.

④ Having more relevant and up to date evidences

⑤ Additional retrieved information beyond the organic search results provides more further gains

⑥ Increasing the Number of Retrieved evidences

⑦ Verbose Demonstration Improve on Complex Questions but also increase

Hall u uhatln.