Will It Still Be True Tomorrow?

Multilingual Evergreen Question Classification to Improve Trustworthy QA

Sergey Pletenev*[1,2], Maria Marina*[2,1], Nikolay Ivanov[1], Daria Galimzianova[3,4], Nikita Krayko[3], Mikhail Salnikov[2,1], Vasily Konovalov[2,1,5], Alexander Panchenko[1,2], Viktor Moskvoretskii[6]

[1]Skoltech [2]AIRI [3]MWS AI [4]MBZUAI [5]MIPT [6]EPFL

Motivation

Large Language Models (LLMs) frequently hallucinate in question answering, particularly when questions are *temporally sensitive*.

Definition

We identify question **temporality** as a critical yet underexplored factor:

- 🌲 **Evergreen:** answers remain stable over time
- 🍂 **Mutable:** answers change, requiring up-to-date information

🌲 **EVERGREEN:** "What is the freezing point of water?"
→ *0°C (stable scientific fact)*

🍂 **MUTABLE:** "Who is the current UK Prime Minister?"
→ *Changes with elections*

Classification model

We experimented with multilingual versions of BERT, DeBERTaV3, and E5. The best performance was achieved using the E5-Large model, which we refer to as our classifier **EverGreen-E5 (EG-E5)**

Model	Type	Russian	English	Chinese	AVG
EG-E5	Trained	**0.910**	**0.913**	**0.897**	**0.906**
LLaMA 3.1-70B	Few-shot	0.889	0.879	0.873	0.875
Qwen 2.5-32B	Few-shot	0.882	0.885	0.872	0.874
Gemma 2-27B	Few-shot	0.830	0.878	0.826	0.838
Deberta v3 base	Trained	0.836	0.900	0.831	0.836
EG-E5 Small	Trained	0.821	0.822	0.817	0.815
GPT-4.1	Few-shot	0.806	0.794	0.809	0.807
UAR	Trained	0.635	0.599	0.731	0.696

Insight

EG-E5 significantly outperforms all few-shot LLMs (including 70B+ parameter models) and prior specialized methods, while remaining lightweight and efficient.

Dataset

We construct a QA dataset consisting of real user queries sourced from an AI chat assistant, each labeled as either 🌲 **Evergreen** or 🍂 **Mutable**. All questions are factual in nature and were manually validated.

- 🌱 **Multilingual** (EN, RU, FR, DE, HE, AR, ZH)
- 🌱 All texts translated by GPT 4.1 and **verified by trained linguists**
- 🌱 **Real queries** from AI assistant with **additional pre-processing**
- 🌱 4,757 per languages making **33,299 in total** for train and test

Dataset	Size	Multilingual	Train	Both classes
TimeQA	∼40k	✗	✗	✗
MuLan	∼246k	✗	✗	✓
FreshQA	600	✗	✗	✓
TAQA	∼20k	✗	✓	✗
EverGreenQA	**33,299**	✓	✓	✓

Table 1. Comparison with existing datasets

Applications

Dataset Flitering

Dataset	Mutable	Dataset	Mutable
NQ	18%	HotpotQA	10%
TriviaQA	6%	MuSiQue	17%
SQuAD	12%	2wikiQA	0.1%

Insight

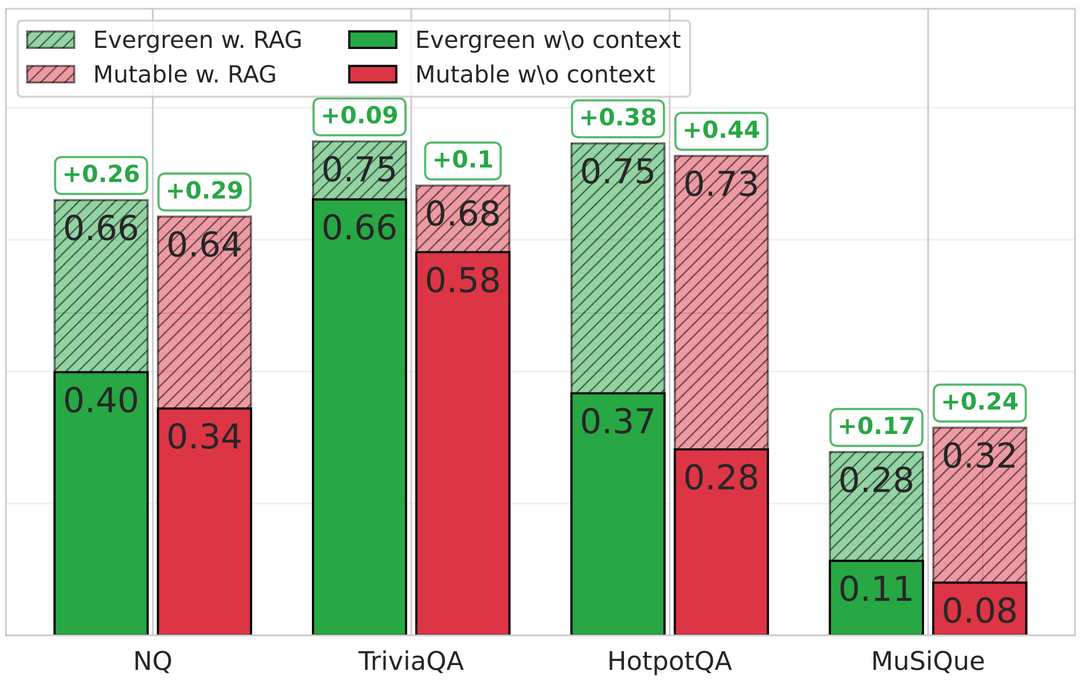
Popular QA benchmarks contain up to **18% mutable questions** with outdated answers, which may unfairly penalize modern LLMs.

Self-Knowledge Estimation

- 🌱 **Evergreen** probability **consistently improves self-knowledge** estimation and calibration over basic uncertainty metrics in 16/18 settings.
- 🌱 If a question is classified as **evergreen, models are much more likely to know the answer.**

RAG gain with EverGreen

- 🌱 A higher accuracy gain indicates RAG is more beneficial for **mutable questions**.



Explaining GPT's "web search"

- 🌱 Evergreen-ness is **strongest predictor of GPT-4o's retrieval behavior.**
- 🌱 More than **twice as informative as uncertainty metrics.**

Predictor	Correlation
Ground Truth	**0.77**
EG-E5	0.66
LLMs 8B	0.20-0.34
LLMs >24B	0.17-0.35

Test a demo!

Conclusion

Question temporality is crucial for trustworthy QA. **EverGreenQA** enables fair evaluation, and **EG-E5** provides efficient evergreen classification. Our work demonstrates practical applications in *self-knowledge estimation*, *dataset curation*, and *explaining LLM retrieval behavior*.

Code and Data

S.Pletenev@skol.tech Maria.Marina@skol.tech A.Panchenko@skol.tech