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**[2,5], **Alexander Panchenko**[1,2], **Viktor Moskvoretskii**[6,**]
[1]Skoltech, [2]AIRI, [3]MWS AI, [4]MBZUAI, [5]MIPT
[6]School of Computer and Communication Sciences, EPFL
{S.Pletenev, Maria.Marina, A.Panchenko}@skol.tech

Abstract

Large Language Models (LLMs) often hallucinate in question answering (QA) tasks. A key yet underexplored factor contributing to this is the temporality of questions – whether they are `evergreen` (answers remain stable over time) or `mutable` (answers change). In this work, we introduce **EverGreenQA**, the first multilingual QA dataset with evergreen labels, supporting both evaluation and training. Using **EverGreenQA**, we benchmark 12 modern LLMs to assess whether they encode question temporality explicitly (via verbalized judgments) or implicitly (via uncertainty signals). We also train **EG-E5**, a lightweight multilingual classifier that achieves SoTA performance on this task. Finally, we demonstrate the practical utility of evergreen classification across three applications: improving self-knowledge estimation, filtering QA datasets, and explaining GPT-4o's retrieval behavior.

1 Introduction

Large language models (LLMs) often struggle with question answering (QA) due to hallucinated answers (Huang et al., 2025). To improve trustworthiness, recent research has focused on estimating LLMs' *self-knowledge* – their ability to recognize what they do and do not know (Yin et al., 2023; Moskvoretskii et al., 2025) – and on integrating up-to-date external information through Retrieval-Augmented Generation (RAG) (Su et al., 2024; Jeong et al., 2024; Trivedi et al., 2023; Belikova et al., 2024).

A particularly important but underexplored factor affecting question difficulty is whether a question is `evergreen` or `mutable` (Wei et al., 2024) – that is, whether its correct answer remains stable over time, as illustrated in Figure 1. Mutable questions are especially challenging because they often require access to up-to-date information, which may be missing from a model's fixed, parametric knowledge.

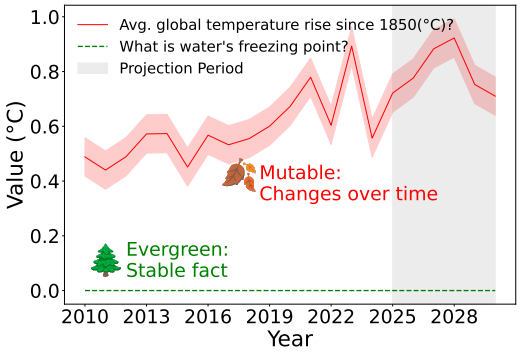

Figure 1: Some questions have answers that stay the same (**evergreen**), like facts of nature. Others have answers that change over time and will change in the future (**mutable**), like global trends or statistics.

Despite its practical importance, evergreen-ness remains an underexplored factor in evaluating and improving LLM behavior. Most existing studies are limited to small-scale, English-only datasets and focus primarily on QA accuracy, rarely examining its broader implications (Vu et al., 2024; Cheng et al., 2024). As a result, the role of question evergreen-ness in shaping LLM reliability and interpretability remains largely unexamined.

To address this gap, we conduct a comprehensive study of question evergreen-ness and its practical applications. We introduce **EverGreenQA** – the first multilingual human-curated evergreen-aware QA dataset, which includes a train–test split suitable for model training. Using **EverGreenQA**, we evaluate 12 modern LLMs to determine whether they encode temporal knowledge explicitly (through direct prompting) or implicitly (via uncertainty-based signals). Further, we develop **EG-E5** – a lightweight SoTA classifier trained to identify evergreen questions.

* Equal contribution.
** Work has been done while at Skoltech.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8615–8632
November 4-9, 2025

	EverGreenQA (our work)	**TimeQA** (Chen et al., 2021)	**MuLan** (Fierro et al., 2024)	**FreshQA** (Vu et al., 2024)	**TAQA** (Zhao et al., 2024)
Both EG and mutable questions	✓	✗	✓	✓	✗
Train-Test split	✓	✓	✗	✗	✓
Human-Evaluated	✓	✓	✗	✓	✗
Multilinguality	✓	✗	✗	✗	✗
Overall size	33,299	~40k	~246k	600	~20k

Table 1: Comparison of EverGreenQA to other time-sensitive datasets.

We demonstrate the usefulness of **EG-E5** in several downstream tasks: (1) improving self-knowledge estimation, (2) curating QA datasets to support fairer evaluation, and (3) effectively explaining GPT-4o's black-box retrieval behavior.

Our contributions and findings are as follows:

1. We construct **EverGreenQA** – the first multilingual dataset for question evergreen-ness classification, covering 7 languages with 4,757 examples per language, resulting in a total of 33,299 examples.
2. We conduct the first comprehensive evaluation of question evergreen knowledge in LLMs, assessing 12 models using both explicit signals (via prompting) and implicit signals (via uncertainty estimation).
3. We develop **EG-E5** – a multilingual lightweight classifier for identifying evergreen questions, which serves as SoTA approach for question evergreen-ness classification while remaining suitable for low-compute settings.
4. We demonstrate the utility of **EG-E5** across three applications: (1) improving self-knowledge estimation, (2) curating QA datasets for fairer evaluation, and (3) effectively explaining GPT-4o's retrieval behavior.

We release the model and data for further usage. [1]

2 Related Work

Reasoning about time remains a fundamental challenge in question answering (QA) tasks, as temporal dynamics often complicate both the interpretation of questions and the retrieval of accurate answers. Working with time in QA tasks has been improved thanks to datasets like *TimeQA* (Chen et al., 2021), which has 20,000 question-answer pairs requiring temporal reasoning. While helpful, it only addressed simple reasoning. *SituatedQA* (Zhang and Choi, 2021) showed the importance of context by situating questions in time and place. *StreamingQA* highlighted the need for temporal adaptation, revealing LLMs' difficulty tracking changing facts (Liska et al., 2022). *TemporalAlignmentQA (TAQA)* (Zhao et al., 2024) further enhances possibility for temporal alignment by providing 20K time-sensitive questions and their answers for each year from 2000 to 2023. *MuLan* (Fierro et al., 2024) differentiated questions by change rate and fact type, respectively. Most recently, *FreshQA* (Vu et al., 2024) introduced a benchmark focused on freshness-sensitive information, further illustrating LLMs' limitations in handling temporally dynamic knowledge. These studies indicate a need for specialized temporal reasoning (Fierro et al., 2024). The comparison of datasets is presented in Table 1.

Retrieval-Augmented Generation (RAG), such as DRAGIN (Su et al., 2024), IRCoT (Trivedi et al., 2023) or Rowen (Ding et al., 2024) was used to solve the problem of time sensitive QA, addressed this through dynamic retrieval decisions, but showed limited results. Alternatively, similar problems can be solved by trying to use structured data, such as knowledge graphs (Salnikov et al., 2025; Lysyuk et al., 2024).

Dynamic retrieval decisions required self-knowledge estimation. In other words, before QA systems can be trusted, they need to know what they don't know. Often, LLMs struggle to identify questions they can't answer (Yin et al., 2023), but using self-knowledge can reduce mistakes in tasks that need a lot of knowledge (Wang et al., 2023; Moskvoretskii et al., 2025).

While retrieval-based methods address temporal knowledge gaps externally, another direction is to update the internal knowledge of LLMs. Updating internal knowledge in LLMs is computationally expensive, as retraining or editing models often requires substantial resources and cannot be performed daily or hourly in practice. Tech-

[1] https://github.com/s-nlp/Evergreen

niques like *LLM Surgery* (Veldanda et al., 2024) and parameter-efficient fine-tuning (Ge et al., 2024; Pletenev et al., 2025) have attempted to make such updates more practical, but still face issues with large-scale changes or factual hallucinations.

3 EverGreenQA & EG-E5

Dataset Collection. Constructing a dataset with both evergreen and mutable questions presents substantial complexity. Deciding whether a fact is temporally stable or subject to change often requires domain expertise, extensive manual verification, and careful distinction of edge cases. Many questions superficially appear evergreen but are actually mutable upon deeper examination, as even "stable" domains like geography or astronomy periodically undergo re-definition or new discoveries. We construct a QA dataset consisting of real user queries sourced from an AI chat assistant, each labeled as either `evergreen` or `mutable`, along with corresponding golden answers. All questions are factual in nature and were manually validated over multiple iterations of internal alpha testing to ensure diversity and reduce topic bias. The labels and golden answers were assigned by a team of trained linguists, who manually wrote the answers from scratch based on retrieved information. Due to the fact that in the initial dataset most of the questions were `mutable` and to avoid bias in the training data, we also generated 1,449 synthetic examples for the `evergreen` class only. This additional dataset was similarly validated by linguists. The final dataset contains 4,757 questions per language, with 3,487 used for training and 1,270 reserved for testing, resulting in a total of 33,299 questions across 7 languages. Details of dataset collection and labeling are presented in Appendix H.

Dataset Translation. We perform translations from Russian to English and from English to the target languages using GPT-4.1, following prior work that demonstrated its strong performance across a wide range of languages, including low-resource ones (Vayani et al., 2025; Chan and Tang, 2024; Simonsen and Einarsson, 2024; Raunak et al., 2023; Yan et al., 2024). The full translation prompt is provided in Appendix C.

Dataset Validation. We performed two complementary validation checks:

Human Evaluation. We recruited human evaluators for each target language, all of whom are either native speakers or possess advanced proficiency (B2–C1 level). We randomly sampled 100 questions from the test set (50 mutable, 50 evergreen) for each language; additionally, for Hebrew we manually verified 200 further items (300 in total). No errors were found in the translations for English, Hebrew, German, or Arabic, while Chinese exhibited only two minor inaccuracies. Validation assessor instruction is provided in Appendix D.

Automatic Evaluation. We further assess the quality of the multilingual data through downstream consistency: our classifier converges reliably and improves QA performance (Table 4, Table 6). We also observe strong generalization, as retraining solely on our dataset and evaluating on FreshQA yields an F1 score of 0.84 (Appendix I).

Question complexity. We additionally verified that evergreen and mutable questions are balanced in terms of complexity, ensuring that observed differences cannot be attributed to question simplicity (details in Appendix F).

EG-E5 Training. For training and testing, we used our multilingual dataset. For validation, we employed the dev and test splits from FreshQA (Vu et al., 2024), merging the `fast-changing` and `slow-changing` classes into `mutable` label. To align with our multilingual setting, the FreshQA data was translated into all target languages.

We experimented with multilingual versions of BERT (Devlin et al., 2019), DeBERTaV3 (He et al., 2023), and E5 (Wang et al., 2024) as encoders. The best performance was achieved using the E5-Large model, which we refer to as our classifier **EverGreen-E5 (EG-E5)**. Hyperparameter details and ablation results are provided in Appendix B.

4 Are LLMs Aware of Evergreenness?

In this section, we evaluate whether modern LLMs can reliably identify whether a given question is evergreen. We test 12 LLMs spanning diverse architectures, with full details provided in Appendix B.

4.1 Verbalized Evergreen Awareness

To assess whether LLMs are capable of explicitly recognizing evergreen questions, we prompt each model to provide a binary Yes/No answer.

We additionally include two specifically trained methods: **UAR** (Cheng et al., 2024): a previously proposed LLaMA2-13b fine-tuned to classify evergreen questions, and **MULAN** (Fierro et al., 2024)

Model	Russian	English	French	German	Hebrew	Arabic	Chinese	AVG
Few-Shot Verbalized Classification								
LLaMA 3.1-8B-it	0.677	0.699	0.686	0.677	0.667	0.659	0.652	0.674
LLaMA 3.1-70B-it	0.889	0.879	0.895	0.87	0.874	0.829	0.873	0.875
Qwen 2.5 7B-it	0.782	0.789	0.786	0.794	0.692	0.711	0.774	0.761
Qwen 2.5 32B-it	0.882	0.885	0.875	0.883	0.862	0.862	0.872	0.874
Qwen 2.5 72B-it	0.806	0.815	0.802	0.805	0.781	0.758	0.768	0.791
Phi-3 medium 4k-it	0.556	0.577	0.498	0.473	0.499	0.498	0.420	0.503
Phi-3 medium 128k-it	0.415	0.489	0.342	0.335	0.385	0.304	0.289	0.366
Gemma 2-9B-it	0.755	0.728	0.694	0.723	0.740	0.711	0.746	0.728
Gemma 2-27B-it	0.830	0.878	0.836	0.827	0.838	0.831	0.826	0.838
Mistral 7B-it-v0.3	0.736	0.722	0.726	0.729	0.670	0.666	0.731	0.711
Mistral Small-24B-it-2501	0.827	0.739	0.768	0.789	0.847	0.834	0.839	0.806
GPT-4.1	0.806	0.794	0.816	0.813	0.803	0.811	0.809	0.807
Trainable Methods								
UAR (Original) (Cheng et al., 2024)	0.550	0.500	0.510	0.600	0.670	0.710	0.710	0.490
UAR (EverGreenQA Data)	0.635	0.599	0.721	0.711	0.698	0.751	0.731	0.696
MULAN (Fierro et al., 2024)	0.340	0.345	0.442	0.379	0.322	0.220	0.279	0.340
EG-E5 (our)	**0.910**	**0.913**	**0.909**	**0.910**	**0.904**	**0.900**	**0.897**	**0.906**

Table 2: Comparison of verbalized LLM predictions and trainable classifiers on the test part of evergreen classification task. Reported scores are weighted F1. A random baseline achieves 0.637. LLMs were prompted with 10-shot examples. The best scores are shown in **bold**. UAR is reported from original paper and trained on our dataset.

classification based on mutable and evergreen samples from Wikidata.

Results. Table 2 shows that our proposed classifier, **EG-E5**, achieves the highest performance across all languages, significantly outperforming both general-purpose and specifically trained LLMs. Among the LLMs, LLaMA 3.1 70B and Qwen 2.5 32B are the strongest, with GPT-4.1 lagging a bit behind.

We observe some variations in language performance, but no clear performance gap, even for non-Latin languages (e.g., Arabic, Chinese, Russian).

Baseline methods UAR and MULAN perform substantially worse than both LLMs and EG-E5, likely due to their oversimplified assumptions regarding the evergreen nature of QA datasets. Even when UAR is trained with our dataset, we observe the underperformance, stemming from less representational power of embeddings.

Takeaway

EG-E5 outperforms few-shot LLMs and prior methods, whose weaker results stem from unrealistic assumptions in their training data.

4.2 Internal Evergreen Awareness

We next assess whether LLMs implicitly encode information about question evergreen-ness through their uncertainty estimates using a balanced subset of sampled 400 questions from our test set – 200 labeled as `evergreen` and 200 as `mutable`.

We select two widely adopted uncertainty measures that show strong performance (Vashurin et al., 2024; Moskvoretskii et al., 2025).

Perplexity – the inverse probability of the predicted sequence, normalized by its length. For a sequence of tokens $x_1, \ldots, x_T$, it is defined as:

$$\text{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t})\right)$$

Mean Token Entropy – the average entropy of the model's predicted token distribution at each position:

$$\text{Entropy} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{w\in V} p_t(w)\log p_t(w)$$

where $p_t(w)$ is the predicted probability of token w at position t, and V is the vocabulary.

Results. Table 3 shows that most models exhibit only mild correlations between uncertainty and evergreen-ness, with Mistral 7B and Qwen 2.5 32B achieving the strongest signals.

We also observe a weak trend suggesting that larger models correlate more strongly with evergreen-ness, possibly indicating a greater internal reliance on temporal cues. Neither perplexity nor entropy consistently outperforms the other. Overall, uncertainty signals capture some temporal information, but are noticeably weaker than explicit verbalized judgments. Additional analysis is provided in Appendix G.

Model	Perplexity	Mean Token Entropy
Gemma 2-9B-it	0.23	0.27
Gemma 2-27B-it	0.26	0.29
LLaMA 3.1-8B-it	0.33	0.33
LLaMA 3.1-70B-it	0.20	0.21
Mistral 7B-it-v0.3	0.33	0.35
Mistral Small-24B-it-2501	0.34	0.32
Phi-3 medium 4k-it	0.23	0.17
Phi-3 medium 128k-it	0.27	0.32
Qwen 2.5 7B-it	0.25	0.25
Qwen 2.5 32B-it	0.33	0.34
Qwen 2.5 72B-it	0.29	0.31

Table 3: Pearson correlation between evergreen-ness and model uncertainty: Perplexity and Mean Token Entropy. All coefficients are statistically significant ($p<0.05$). See Appendix G for additional analysis.

Takeaway

Uncertainty metrics encode weak and inconsistent signals of evergreen-ness, with slightly stronger trends in larger models.

5 Enhancing Self-Knowledge

In this section, we evaluate whether incorporating knowledge about question evergreen-ness improves the estimation of *self-knowledge* – a model's ability to recognize the boundaries of its own knowledge and determine when it can or cannot answer a given question (Moskvoretskii et al., 2025; Yin et al., 2023). This capability is considered a key factor in improving the trustworthiness of LLMs.

5.1 Task Formulation

We frame self-knowledge estimation as a binary classification task, where the target label $y \in \{0, 1\}$ reflects whether the model's answer to a given input x is factually correct. Each method under evaluation assigns a real-valued self-knowledge score $f(x) \in \mathbb{R}$ to the input.

5.2 Methods

We evaluate this setup using LLaMA3.1-8B-Instruct with five widely adopted and high-performing uncertainty estimators, selected to represent different families of uncertainty quantification methods – including logit-based and consistency-based approaches:

Max Token Entropy: Evaluates uncertainty by computing token-level entropies and taking the maximum value across the sequence as the final score (Fomicheva et al., 2020).

Mean Token Entropy: Similar to the above, but aggregates across the sequence by averaging token-level entropy values (Fomicheva et al., 2020).

Lexical Similarity: Estimates uncertainty by calculating the average lexical overlap among multiple model responses, serving as a proxy for output consistency (Fomicheva et al., 2020).

SAR: Combines entropy with relevance weighting by amplifying the contribution of semantically important tokens, summing the adjusted entropy values over the sequence (Duan et al., 2024).

EigValLaplacian: Constructs a similarity graph over sampled responses and computes the sum of eigenvalues of its Laplacian matrix to quantify response diversity (Lin et al., 2024).

For each method, we evaluate the effect of incorporating the predicted probability of a question being evergreen, obtained from our trained evergreen classifier.

To obtain the final self-knowledge classifier $f(x)$, we train a standard machine learning model on the training set, using the uncertainty estimation metrics as input features. When applicable, we also include the predicted evergreen probability as an additional feature. The full training procedure is detailed in Appendix E.

5.3 Evaluation

We evaluate performance using standard metrics widely adopted in recent literature on uncertainty estimation (Fadeeva et al., 2024; Vashurin et al., 2024; Vazhentsev et al., 2025).

AUROC measures how well the model distinguishes between correct and incorrect answers based on the self-knowledge score $f(x)$. Higher values indicate stronger separability.

AUPRC quantifies the trade-off between precision and recall across different decision thresholds. It is particularly informative when dealing with imbalanced datasets.

Prediction Rejection Ratio (PRR) measures how well uncertainty scores align with answer quality. It simulates rejecting the most uncertain responses and tracks how average quality improves. Higher PRR indicates better calibration between uncertainty and actual answer correctness. We use In-Accuracy as main QA metric.

5.4 Datasets

We evaluate our methods on 6 QA datasets covering both single-hop and multi-hop reasoning. The single-hop datasets include SQuAD

Method	NQ			SQuAD			TriviaQA			2WikiMultihopQA			HotpotQA			MuSiQue		
	AUROC	AUPRC	PRR	AUROC	AUPRC	PRR	AUROC	AUPRC	PRR	AUROC	AUPRC	PRR	AUROC	AUPRC	PRR	AUROC	AUPRC	PRR
Uncertainty Estimation																		
EigValLaplacian	0.56	0.46	0.56	0.48	0.19	0.77	0.70	0.52	0.58	0.61	0.26	0.71	0.64	0.22	0.75	0.57	0.09	0.86
LexicalSimilarity	**0.61**	0.38	0.59	0.64	0.13	0.83	0.65	0.54	0.58	0.55	0.29	0.67	0.68	0.21	0.77	0.58	0.09	0.85
MaxTokenEnt.	**0.61**	0.37	**0.60**	0.58	0.18	0.80	0.70	0.51	0.62	0.59	0.27	0.69	0.67	0.21	0.75	0.64	0.09	0.84
MeanTokenEnt.	0.59	0.42	0.57	0.56	0.19	0.80	0.71	0.50	**0.63**	0.61	0.28	0.70	0.62	0.23	0.73	0.63	0.09	0.81
SAR	**0.61**	0.39	0.59	0.67	0.12	0.84	**0.72**	0.51	0.60	0.60	0.27	0.69	0.69	0.21	**0.78**	0.64	0.08	0.83
Uncertainty Estimation + Evergreen																		
EigValLaplacian+EG	0.56	0.40	0.57	0.49	0.19	0.77	0.70	0.51	0.56	0.54	**0.52**	0.64	0.65	0.21	0.75	0.50	**0.12**	0.84
LexicalSimilarity+EG	0.59	0.40	0.59	0.65	0.13	0.83	0.68	0.52	**0.63**	0.61	0.26	0.71	0.68	0.21	0.76	0.61	0.10	0.86
MaxTokenEnt.+EG	0.56	0.42	0.59	0.68	0.12	0.85	0.71	0.51	**0.63**	**0.63**	0.25	**0.72**	0.67	0.21	0.75	0.55	0.11	0.84
MeanTokenEnt.+EG	0.59	0.39	0.59	**0.70**	0.12	**0.86**	**0.72**	0.50	0.62	0.61	0.26	0.71	0.63	0.22	0.75	0.64	0.08	0.85
SAR+EG	0.58	0.41	0.57	**0.70**	0.12	0.85	0.66	0.54	0.46	0.62	0.43	0.68	**0.70**	0.21	**0.78**	**0.67**	0.11	**0.87**
Evergreen																		
EG	0.50	**0.72**	0.52	0.52	**0.20**	0.79	0.47	**0.65**	0.62	0.49	0.31	0.65	0.51	**0.28**	0.68	0.50	0.10	**0.87**

Table 4: Self-knowledge identification performance. We report classification quality using AUROC and AUPRC, and calibration efficiency using PRR. EG stand for Evergreen probability. Higher values indicate better performance. The best scores for each metric are shown in **bold**.

v1.1 (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017), while the multi-hop datasets include MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and 2WikiMultiHopQA (Ho et al., 2020).

Following Trivedi et al. (2023); Jeong et al. (2024), we use a subset of 500 questions from the original splits of each dataset to ensure consistency and comparability.

5.5 Results

As shown in Table 4, evergreen probability is a strong signal for improving self-knowledge identification. In 16 out of 18 evaluations, the best results are achieved either by the evergreen feature alone or by combining it with an uncertainty estimation method. Moreover, it is able to improve calibration (PRR) which depends on QA accuracy, making it highly valuable for real-world applications.

Notably, the evergreen feature alone performs exceptionally well on AUPRC, achieving the top score on 4 datasets. This suggests that evergreenness is a powerful indicator of when a model possesses reliable knowledge.

However, we also observe a consistent pattern: evergreen scores high on AUPRC but relatively low on AUROC. This indicates that while the feature is highly effective at identifying when the model knows the answer, it is less reliable at recognizing when the model does not (weaker true negative discrimination). In other words, if a question is evergreen, the model is likely to answer it correctly – but if a question is not evergreen, the outcome is harder to predict.

Takeaway

Evergreen probability consistently improves self-knowledge estimation and calibration, achieving top results in 16 out of 18 settings.

6 Filtering QA with Evergreen

In this section, we demonstrate that evergreen classification is valuable for filtering QA datasets, enabling fairer evaluation by excluding mutable questions. We use the same model setting as in Self-Knowledge Section 5.

QA datasets should ideally consist only of evergreen questions, emphasized in SimpleQA (Wei et al., 2024). To achieve this, SimpleQA relied on human annotators to assess evergreen-ness. In contrast, **EG-E5** enables automated dataset curation, eliminating the need for manual annotation and facilitating the scalable construction of large QA corpora.

6.1 Popular QA Datasets Analysis

Mutable questions pose a serious challenge for fair QA evaluation: outdated gold answers can make correct responses from modern LLMs appear wrong, especially when models are evaluated at different times.

Examples. Table 5 highlights such mutable examples across six datasets (Section 5.4), showing answers that, as of 2025, diverge from the original references. These include both simple and complex queries - even from recently released datasets like MuSiQue (Trivedi et al., 2022). The nature of change varies: some are predictable (e.g., Olympic host cities, population figures), some occasional

Dataset	Dataset release	Non-EG question	Reference answer	Answer in 2025
		Expected changes (routine, scheduled)		
NQ	2015	what city is the next winter olympics in	Beijing	Milan
MuSiQue	2022	Who is the mayor presiding now where Merrill Elam was born?	Lance Bottoms	Andre Dickens
SQuAD	2020	How many teams are in the Greek Super League?	18	14
HotpotQA	2018	Yau Ma Tei North is a district of a city with how many citizens?	7.2 million	7.4 million
MuSiQue	2022	According to QS World University Rankings, where does the college that Ibrahim Shihata attended rank?	551-600	350
		Occasional changes (updates over time, but less regular)		
HotpotQA	2018	Edoardo Soleri is playing on loan from which Italian football club?	A.S. Roma	Spezia
2WikiMultihopQA	2020	Where does Karin Stoltenberg's husband work at?	United Nations	He has died
2WikiMultihopQA	2020	Who is the spouse of the performer of song Les Rois Du Monde?	Joy Esther	Emily Surde
TriviaQA	2017	What is the name of the current Attorney General for England and Wales?	Dominick Grieve	Richard Hermer
NQ	2015	who is the current minister for environment forest and climate change in india	Dr. Harsh Vardhan	Bhupender Yadav
		Less predictable changes (complex sociopolitical shifts)		
SQuAD	2020	What is the largest economy in Africa?	Nigeria	South Africa
TriviaQA	2017	Who is fifth in line to the throne?	Princess Beatrice	Prince Harry

Table 5: Examples of non-evergreen questions from popular QA datasets, showing discrepancies between original gold answers and updated answers in 2025. Questions are categorized by the nature of the change: expected, occasional, and less predictable.

Dataset	0-Shot		RAG		Δ EG-Mut	Mut RAG	Mut, %
	EG	Mut	EG	Mut	0-shot, %	Gain, %	
NQ	0.399	0.344	0.660	0.635	16	10	18
TriviaQA	0.661	0.581	0.749	0.682	14	13	6
SQuAD	0.171	0.168	0.627	0.598	2	-6	12
HotpotQA	0.367	0.282	0.746	0.727	30	14	10
MuSiQue	0.113	0.080	0.278	0.315	41	30	17
2wikiMultihopQA	0.448	0.342	0.644	0.457	31	-70	0.1

Table 6: Performance comparison between evergreen (EG) and mutable (Mut) questions under 0-shot (no context) and RAG (with context) settings. We report absolute in-accuracies, the relative gap between evergreen and mutable questions (Δ EG–Mutable) under 0-shot, and the relative RAG gain on mutable questions. A higher mutable gain indicates RAG is more beneficial for time-sensitive queries. The last column shows the proportion of mutable questions in each dataset. Gray row indicates limited applicability due to extremely low mutable sample count.

(e.g., job titles or spouses), and others unexpected (e.g., monarchs, GDP rankings).

Statistics. Table 6 shows that mutable questions remain common, reaching 18% in NQ and averaging 10% across datasets. This challenges the widespread assumption that QA benchmarks are temporally stable, and raises concerns about evaluation fairness. To ensure reliability, mutable questions should be filtered out, or alternatively, live benchmarks like RealTimeQA (Kasai et al., 2023) should be maintained-though they are costly to sustain.

Incorrect Assumptions. UAR (Cheng et al., 2024) has implicitly assumed dataset evergreenness and MULAN (Fierro et al., 2024) treat many questions as immutable, yet some relations (e.g., Wikidata's P190, "sister cities") can in fact change. This mismatch may help explain the limited real-world effectiveness of such methods when faced with temporal drift.

Takeaway

QA benchmarks include mutable questions, undermining fair evaluation. Filtering for evergreen questions is essential for reliable assessment.

6.2 Filtered QA Performance

Zero-Shot Performance. As shown in Table 6, model accuracy is consistently higher on evergreen questions, with relative differences reaching up to 40% on complex tasks. This aligns with expectations, as mutable questions often require up-to-date information beyond the model's static knowledge.

RAG Benefits. We show that models generally benefit more from RAG with gold contexts when answering mutable questions, with relative gains

Misclassification reason	Example questions
False Positives (non-evergreen, but classified as evergreen)	
Superlatives assumed to be static facts	· What is the biggest star in the sky? · Which tea is the healthiest? · What is the most popular social network in the world?
Biographical/life data on alive people treated as static	· In which movies has Simu Liu acted? · How many works has Stephen King written?
False Negatives (evergreen, but classified as non-evergreen)	
Superlatives treated as time-sensitive or trend-based	· What is the oldest currency? · The rarest element in the periodic table. · How long did the shortest war in history last?
Biological and geographical facts wrongly assumed to change frequently	· What is the area of Liechtenstein? · Which animal has the highest blood pressure? · How many species of elephants currently live on the planet?

Table 7: Error Analysis of EG-E5 Classifier: breakdown of misclassification patterns.

Model	ChatGPT
Gemma 2-9B-it	0.26
Gemma 2-27B-it	0.30
LLaMA 3.1-8B-it	0.29
LLaMA 3.1-70B-it	0.25
Mistral 7B-it-v0.3	0.34
Mistral Small-24B-it-2501	0.33
Phi-3 medium 4k-it	0.20
Phi-3 medium 128k-it	0.29
Qwen 2.5 7B-it	0.28
Qwen 2.5 32B-it	0.36
Qwen 2.5 72B-it	0.35
EG-E5	0.66
EverGreen	0.77

Table 8: Correlation of ChatGPT with UC and EG. All results are significant (p-value < 0.05). EverGreen denotes ground true labels in the selected dataset part.

reaching up to 30%. However, this effect diminishes in datasets with few mutable examples.

7 Explaining GPT-4o Retrieval

GPT-4o autonomously decides when to invoke its retrieval system using internal, black-box criteria. We find that question evergreen-ness is the strongest predictor of this behavior, suggesting that GPT-4o's use of external search is closely linked to the temporal nature of the input.

We use the same subset, as in Section 4.2 – and queried GPT-4o via its web interface,[2] recording whether it triggered a retrieval call.

In addition to evergreen labels, we evaluated several uncertainty-based signals from Section 4.2 and **EG-E5** to assess their correlation with GPT-4o's retrieval decisions.

[2]All experiments with GPT-4o were performed in May 2025 using the publicly available web interface.

As shown in Table 8, evergreen-ness and **EG-E5** predictions are substantially stronger predictors than any uncertainty-based signal – more than twice as informative. This suggests that GPT-4o may internally model question temporality or is guided by a retrieval policy highly sensitive to it.

Takeaway

Evergreen-ness is the strongest predictor of GPT-4o's retrieval behavior, suggesting that retrieval is closely tied to temporality.

8 Error Analysis

We selected a test part from our EverGreenQA dataset and conducted a qualitative analysis of the errors made by the EG-E5 classifier. Table 7 presents examples of false positives and false negatives, grouped by cause. Notably, the classifier shows high uncertainty with superlatives — sometimes flagging them as volatile, and other times misinterpreting trend-sensitive phrases like 'most,' 'biggest,' or 'healthiest' as universally fixed. Other errors include misclassifying achievements of living people as dead and incorrectly treating stable geographical or biological facts as time-sensitive.

Interestingly, there are twice as many false negatives as false positives. This suggests that the classifier is more cautious when deciding whether a question refers to a stable fact. In some cases, external information is crucial. For example, if a person is dead, all questions about them would be evergreen, but the model needs to know whether the person is still alive. Similarly, questions about recent years (e.g., 2023–2024) pose a challenge, as

the model lacks awareness of the current date. In other cases, there is room for improvement in how the model organizes and distinguishes its knowledge. For instance, learning to differentiate between truly stable physical facts (such as the area of Liechtenstein) and more variable ones (like the brightest star in the sky), or between completed historical events (e.g., the French Revolution) and ongoing developments (such as upcoming presidential elections).

Additional examples are provided in Appendix J.

Conclusion

In this study, we explored the concept of evergreenness, whether the answer changes over time. We examined the ability of LLMs to detect it and demonstrated its usefulness across several applications.

To support this investigation, we introduce **EverGreenQA**, a new multilingual dataset comprising 4,757 examples for each of 7 languages, forming 33,299 samples in total. Using this dataset, we benchmark modern LLMs on the task of evergreen question classification and train **EG-E5** – a lightweight classifier that outperforms both LLMs and previously trained methods.

We further analyze whether LLMs implicitly encode evergreen-ness through their uncertainty estimations and find that they encode it weakly, with larger models doing so more consistently. We further enhance existing uncertainty estimators with predicted evergreen probabilities, yielding consistent improvements.

We also show that our evergreen classifier helps curate high-quality QA datasets and supports more reliable and fair evaluations. Finally, we demonstrate that evergreenness is the best predictor of GPT-4o's search behavior, outperforming all other tested factors.

In **future work**, evergreen classifier developed in this work offers significant potential for improving LLM robustness. Future research could integrate classifier into dataset curation processes for pre-training, SFT and especially RLHF stages, automatically filtering out mutable questions that may become outdated and compromise model reliability over time. By addressing the temporal dimension of question answering, this work contributes to the development of more trustworthy and reliable LLM based systems capable of handling dynamic real-world information environments.

Limitations

- While our EverGreenQA dataset is the first multilingual, human-curated benchmark for question temporality, its size remains relatively modest (4,757 examples per language). Nonetheless, it offers high-quality coverage across seven diverse languages and is sufficient to reveal clear trends in model behavior.
- Although we cover 7 languages, the dataset does not span all major language families, and performance in truly low-resource settings remains unexplored. That said, our selection includes both Latin and non-Latin scripts, enabling meaningful multilingual evaluation.
- Our LLM evaluation includes 14 models across a wide range of scales and families, but we primarily focus on representative models from each size tier. Extending to more instruction-tuned or domain-adapted variants could further generalize the findings.
- For uncertainty-based analysis, we focus on five representative metrics. While these are widely used and sufficient to draw strong conclusions, incorporating more recent or task-specific metrics may provide additional insights.
- Our trained evergreen classifier demonstrates strong results, but we perform only limited ablations on its architecture, training procedure, and the use of auxiliary data. Exploring more model variants or transfer learning strategies could further improve robustness.
- Finally, while we demonstrate several practical uses of evergreen classification, we do not explore its potential in tasks such as active learning, answer calibration, or search reranking. We leave these promising directions for future work.

Ethical Considerations

Our work involves the construction and analysis of a multilingual QA dataset, as well as the evaluation of LLM and classifier-based approaches for detecting question temporality. We made a great effort to take into account the following ethical considerations and discuss them to prevent misusage:

All questions in the constructed dataset were sourced from anonymized real-user queries during

internal alpha testing. No personally identifiable information (PII) was collected, stored, or used. All examples are factual in nature and were manually reviewed to ensure compliance with privacy and ethical standards.

Dataset labels and translations were created by trained linguists and multilingual annotators. Annotators were compensated fairly according to local labor regulations. We ensured that the task complexity was reasonable and the working conditions were ethical.

The lightweight classifier and dataset are intended to support research in trustworthy QA and dataset curation. We caution against deploying these tools in high-stakes applications without rigorous domain-specific validation.

Although evergreen classification can help flag outdated or unstable information, it should not be viewed as a substitute for fact verification or timeliness. We explicitly discourage the use of our tools for censorship or exclusion of mutable information inappropriately.

We believe this work contributes to more transparent and interpretable QA systems by introducing temporality as an explicit factor, while taking steps to ensure fairness, privacy, and responsible development.

Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the R.F. in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech №139-10-2025-033.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Julia Belikova, Evegeniy Beliakin, and Vasily Konovalov. 2024. JellyBell at TextGraphs-17 shared task: Fusing large language models with external knowledge for enhanced question answering. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 154–160, Bangkok, Thailand. Association for Computational Linguistics.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Venus Chan and William Ko-Wai Tang. 2024. Gpt for translation: A systematic literature review. *SN Comput. Sci.*, 5(8).

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 17153–17166. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *CoRR*, abs/2402.10612.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,

Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9367–9385. Association for Computational Linguistics.

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024. Mulan: A study of fact mutability in language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 762–771. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Xiou Ge, Ali Mousavi, Edouard Grave, Armand Joulin, Kun Qian, Benjamin Han, Mostafa Arefiyan, and Yunyao Li. 2024. Time sensitive knowledge editing through efficient finetuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 583–593, Bangkok, Thailand. Association for Computational Linguistics.

John T Hancock and Taghi M Khoshgoftaar. 2020. Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):94.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7036–7050. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: what's the answer right now? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models.

In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.

Maria Lysyuk, Mikhail Salnikov, Pavel Braslavski, and Alexander Panchenko. 2024. Konstruktor: A strong baseline for simple knowledge graph question answering. In *Natural Language Processing and Information Systems - 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings, Part II*, volume 14763 of *Lecture Notes in Computer Science*, pages 107–118. Springer.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6355–6384. Association for Computational Linguistics.

Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. 2025. How much knowledge can you pack into a LoRA adapter without harming LLM? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4309–4322, Albuquerque, New Mexico. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Mikhail Salnikov, Andrey Sakhovskiy, Irina Nikishina, Aida Usmanova, Angelie Kraft, Cedric Möller, Debayan Banerjee, Junbo Huang, Longquan Jiang, Rana Abdullah, and 1 others. 2025. Shortpathqa: A dataset for controllable fusion of large language models with knowledge graphs. In *International Conference on Applications of Natural Language to Information Systems*, pages 95–110. Springer.

Annika Simonsen and Hafsteinn Einarsson. 2024. A human perspective on GPT-4 translations: Analysing Faroese to English news and blog text translations. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 24–36, Sheffield, UK. European Association for Machine Translation (EAMT).

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12991–13013. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19565–19575. Computer Vision Foundation / IEEE.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025. Token-level density-based uncertainty quantification methods for

eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2246–2262. Association for Computational Linguistics.

Akshaj Kumar Veldanda, Shi-Xiong Zhang, Anirban Das, Supriyo Chakraborty, Stephen Rawls, Sambit Sahu, and Milind R. Naphade. 2024. LLM surgery: Efficient knowledge unlearning and editing in large language models. *CoRR*, abs/2409.13054.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13697–13720. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10303–10315. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *CoRR*, abs/2411.04368.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *CoRR*, abs/2407.03658.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7371–7387. Association for Computational Linguistics.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15015–15040. Association for Computational Linguistics.

A License and Infrastructure

All experiments were conducted using 1–2 NVIDIA A100 GPUs, totaling approximately 40 GPU-hours. Model usage adhered to their respective licenses: LLaMA 3.1 (Dubey et al., 2024) and Gemma 2 (Rivière et al., 2024) under custom licenses, Phi 3 (Abdin et al., 2024) and E5 under MIT, and Qwen 2.5 (Yang et al., 2024) and Mistrals (Jiang et al., 2023) under Apache 2.0. GPT Models were accessed via API or web-interface.[3] We release our dataset and classifier under the MIT License.

B Evergreen Testing Details

LLM Verbal Parameters. Each example comes with 5-shot for mutable and 5-shot for immutable examples. For llama 3.1 sampling parameters are following: TEMPERATURE=0.7, TOP_P=0.9. For Qwen 2.5: TEMPERATURE=0.6, TOP_P=0.95, TOP_K=20, MIN_P=0

Our Classificator Parameters. All models were trained for 10 epochs with early-stopping and `lr = 4.6e-5, bs = 16`. Additional datasets were not used. We trained one model for all languages. As shown in Table 10 **multilingual-e5-large-instruct** gives best results.

[3] https://openai.com/

Evergreen Verbal Instruction

You are a helpful assistant. You help user to classify the questions based on the temporality. There are two classes: immutable and mutable. Immutable, in which the answer almost never changes. Mutable, in which the answer typically changes over the course of several years or less. Think about each question and in the end answer with Mutable or Immutable starting with 'Classification:'

C Translation Prompt

Translation Validation Instruction

Translate the following English text into French, German, Hebrew, Arabic and Chinese. Provide the translations as a JSON object with keys 'French', 'German', 'Hebrew', 'Arabic', 'Chinese'.

We use GPT 4.1 with TEMPERATURE=0.2 and additional tag "RESPONSE_FORMAT": "JSON_OBJECT"

D Validation Instructions

Translation Validation Instruction

For each translated question, assign a score according to the following criteria:

- **0** – the translation contains errors that *distort the meaning*.
- **1** – the translation contains *minor errors* that do *not affect the overall meaning*.

E Classifier for Self-Knowledge

We explored seven classification models using `scikit-learn` (Buitinck et al., 2013) and `CatBoost` (Hancock and Khoshgoftaar, 2020): Logistic Regression, k-Nearest Neighbors, Multilayer Perceptron, Decision Tree, Random Forest, Gradient Boosting, and CatBoost. All models were trained with standardized features using `StandardScaler`. Hyperparameters were optimized on a validation subset of 100 examples randomly sampled from the training data, and experiments were repeated with three random seeds per dataset to ensure robustness.

For final evaluation, we selected the two best-performing models on the validation set and combined them into a soft-voting ensemble using `VotingClassifier`. Each component model was retrained on the full training set with its tuned hyperparameters.

Hyperparameters grid. *Logistic Regression* : C: [0.01, 0.1, 1], solver: [lbfgs, liblinear], class_weight: [balanced, 0: 1, 1: 1, None], max_iter: [10000, 15000, 20000]

KNN : n_neighbors: [5, 7, 9, 11, 13, 15], metric: [euclidean, manhattan], algorithm: [auto, ball_tree, kd_tree], weights: [uniform, distance]

MLP : hidden_layer_sizes: [(50,), (100,), (50, 50), (100, 50), (100, 100)], activation: [relu, tanh], solver: [adam, sgd], alpha: [0.00001, 0.0001, 0.001, 0.01], learning_rate: [constant, adaptive], early_stopping: True, max_iter: [200, 500]

Decision Tree : max_depth: [3, 5, 7, 10, None], max_features: [0.2, 0.4, sqrt, log2, None], criterion: [gini, entropy], splitter: [best, random]

CatBoosting: iterations: [10, 50, 100, 200], learning_rate: [0.001, 0.01, 0.05], depth: [3, 4, 5, 7, 9], bootstrap_type: [Bayesian, Bernoulli, MVS]

Gradient Boosting: n_estimators: [25, 35, 50], learning_rate: [0.001, 0.01, 0.05], max_depth: [3, 4, 5, 7, 9], max_features: [0.2, 0.4, sqrt, log2, None]

Random Forest: n_estimators: [25, 35, 50], max_depth: [3, 5, 7, 9, 11], max_features: [0.2, 0.4, sqrt, log2, None], bootstrap: [True, False], criterion: [gini, entropy], class_weight: [balanced, 0: 1, 1: 1, None]

F Question Complexity

To ensure that temporal stability is not confounded with question simplicity, we conducted an additional analysis of question complexity. We randomly selected 200 evergreen and 200 mutable questions from the test set and asked GPT-4.1 to classify them as *Simple* (all relevant information explicit) or *Complex* (requiring additional inference). Using few-shot examples from FreshQA, we found that mutable questions were 65.5% simple and 34.5% complex, while evergreen questions were 71.5% simple and 28.5% complex. Since both categories are dominated by simple questions with comparable distributions, the classifier is unlikely to rely solely on question simplicity as a proxy for temporal stability. This analysis further supports the robustness of our findings.

Model	Perplexity	Mean Token Entropy
Gemma 2 9B	0.014	0.070
Gemma 2 27B	0.070	0.013
LLaMA 3.1 8B	0.054	0.066
LLaMA 3.1 70B	0.028	0.021
Mistral 7B	0.016	0.012
Mistral 24B	0.046	0.026
Phi-3-mini 4k	0.032	0.016
Phi-3-mini 128k	0.137	0.073
Qwen 2.5 7B	0.025	0.016
Qwen 2.5 32B	0.020	0.027
Qwen 2.5 72B	0.031	0.029

Table 9: McFadden's pseudo-R^2 scores from logistic regression models trained to predict evergreen probability from two uncertainty metrics: perplexity and mean token entropy.

G Predictive Analysis of Uncertainty for Temporality

Table 9 reports McFadden's pseudo-R^2 values from logistic regression models trained to predict evergreen-ness based on two uncertainty metrics: perplexity and mean token entropy.

Across most models, the pseudo-R^2 scores remain below 0.07, indicating that uncertainty alone provides limited predictive power for evergreen classification. The only notable exception is Phi-3-medium (128k), which achieves the highest scores–0.137 (perplexity)–suggesting that longer context training may improve temporal uncertainty encoding, however still very limited.

We observe no consistent advantage of one uncertainty metric over the other. Similarly, model size does not correlate clearly with predictive performance; smaller models sometimes match or outperform their larger counterparts.

The results indicate that uncertainty metrics capture limited signals of temporality, supporting their use as complementary features rather than standalone predictors of evergreen-ness.

H Dataset Collection Details

The team of trained linguists responsible for assigning the evergreen and mutable labels, as well as writing the golden answers, each hold at least a bachelor's degree in linguistics, ensuring a strong foundation in linguistic principles and effective communication. Additionally, each stage of the labeling process was carefully validated through consultation with the team lead, who provided oversight to maintain consistency and accuracy across the dataset. Furthermore, to support diverse applications, all answers were converted into a set of aliases. The procedure for this conversion is detailed in Appendix H.5. The assessors were fairly paid according to local regulations.

H.1 Golden Answers Annotation

Golden answers should be complete and useful for the user.

Examples of good and informative answers: **Question:** Who is considered the founder of physics? **Answer:** Isaac Newton is widely regarded as the founder of physics. **Comment:** The question is asked in the singular form, and according to many sources, Newton is indeed considered the founder of classical physics. Based on logic, online sources, and answers from competing systems, it's clear that Galileo Galilei and René Descartes also made significant contributions. However, since the question refers to a single person and sources support it, Newton is the most accurate and accepted answer in this context.

Question: Who was the President of Italy in the year 2000? **Answer:** Carlo Azeglio Ciampi was an Italian statesman, the 10th President of the Italian Republic, and former Prime Minister of Italy. **Comment:** A quick fact-check (as should be done for all examples in the guidelines) confirms this answer is accurate and complete.

Example of an incomplete or partially useful answer that is not suitable as a golden answer: **Question:** Do spiders have teeth? **Answer:** Yes, spiders have teeth. **Comment:** A fact-check in open sources reveals that this answer is not accurate enough to be considered a golden answer. The correct response would be: "Spiders do not have teeth, but they have chelicerae, which contain ducts from venom glands that secrete digestive enzymes." Sometimes, chelicerae are colloquially referred to as "fangs" or "teeth", but they are not actually teeth. Therefore, the original answer should be revised to meet the standard of a golden answer.

Birthday-related questions: If the question is phrased like *How old is Yann LeCun?*, the answer should include the exact age, not just the date or year of birth.

Open-ended list questions: For questions such as *What are the tallest mountains?*, *Which astronauts are there?*, or *What animals live in Africa?*, a good answer should list at least several correct examples and include a note that this is not an exhaustive list - more exist.

H.2 Evergreen-ness Annotation

The `evergreen` criterion is a nuanced one. Most questions are considered evergreen because they are related to established facts or events. However, there are domains, such as astronomy, where new discoveries occur regularly. For example, the record for the largest known star has changed quite recently.

The definition of this criterion depends on the domain of the question. In most cases, facts that have remained unchanged for 20–30 years are treated as established. Obviously, questions like *Who is the president?* are not considered evergreen due to frequent changes in political leadership.

During annotation, when we encountered ambiguous cases, we often relied on domain-specific common sense. For example, it is fairly obvious that most major geographical discoveries have already been made. It is highly unlikely that a new largest lake or a previously unknown landmass on our planet will be discovered.

As for questions involving dates, events, and notable personalities, the vast majority of these are considered evergreen, it is nearly impossible to imagine a scenario in which the dates of significant historical events or key facts from someone's biography would change.

Mutable questions:

(1) *What year was the last solar eclipse?*

(2) *Which country has the longest railway?*

(3) *What date does Ramadan begin?*

(4) *When is the next Olympics?*

(5) *How old is Mike Tyson?*

Evergreen questions:

(1) *Into which two states was the Roman Empire divided, and when?*

(2) *Who is Messi?*

(3) *Name the years of Paul von Hindenburg's leadership in Germany.*

(4) *Name the largest lakes on our planet.*

(5) *What is the total area of Europe?*

H.3 Linguistic Criteria for Evergreen Questions

- **Referential Stability**. The question refers to facts, events, or relations that are extremely unlikely to change over decades or centuries (fundamental scientific facts, historical events with fixed timelines, or established cultural knowledge).
- **Absence of Temporal Indexicals**. The question does not contain explicit time indicators such as "current", "now", "in [year]" , "recent", "last", "next", or "as of today" (Who discovered oxygen? -> evergreen, What year were the last Olympic Games -> non-evergreen).
- **Static Nominal Phrasing**. Questions use general, context-invariant noun phrases. They avoid use of titles, roles, or superlatives about living persons or entities, which are more likely to shift (What is the chemical symbol of gold? -> evergreen, Which country is the richest one? -> non-evergreen).
- **Independence from Trend or Popularity**. Avoid referencing the "most popular", "biggest", "newest", or similar dynamic superlatives unless the referent is historically persistent and highly unlikely to shift based on new data or public opinion (What is the most popular social network? -> non-evergreen).

H.4 Synthetic Data Generation

To augment our training data, we generated and manually validated 1,449 additional question–answer pairs using GPT-4.1. Duplicate questions were filtered out, and common templates – such as "how old is the person" – were rephrased to reduce redundancy. We also followed the FreshQA style to diversify the data: the model generated both evergreen and mutable examples, with mutable questions further categorized into two subtypes. This approach enhanced the variety and coverage of our training set.

Synthetic Instruction

Can you generate different question-answer pair: slow-changing questions, in which the answer typically changes over the course of several years (up to 10); fast-changing question, in which the answer typically changes within a year or less; never-changing, in which the answer never changes.

H.5 Short-Answer Generation Prompt

Short-Answer Generator Instruction

You are a **short-answer generator**. Given a factual **question** and a complete (possibly long) **answer**, return several *concise, semantically-equivalent* answer variants.

RULES

1. Every variant must be factually correct and answer the question on its own.
2. Keep each variant as short as possible (≈1–5 words) while still unambiguous.
3. Include the most common spellings, abbreviations, numerals ↔ Roman-numeral forms, and the canonical full form.
4. Do **not** add information that is not explicitly in the answer.
5. Return a JSON object **exactly** like:
{ "answers": ["Variant 1", "Variant 2", . . .] }
EXAMPLES
Question: "Who is king of England?"
Answer: "The King of Great Britain – Carl 3 (Charles Philip Arthur George)." → ["Carl 3", "King is Carl 3", "Carl III", "Charles III", "Charles Philip Arthur George"]

Question: "What is the highest mountain in the world?"
Answer: "Mount Everest is the highest mountain above sea level."
→ ["Mount Everest", "Everest", "Mt. Everest"]

Question: "Which element has the chemical symbol 'O'?"
Answer: "The chemical element with symbol O is oxygen." → ["oxygen", "Oxygen", "element O is oxygen"]

Question: "Who wrote the play 'Romeo and Juliet'?"
Answer: "'Romeo and Juliet' was written by William Shakespeare."
→ ["William Shakespeare", "Shakespeare"]

Question: "What is the currency of Japan?"
Answer: "The Japanese currency is the yen."
→ ["yen", "Japanese yen", "JPY"]

You only have to send **one message** per call.

We query GPT-4o with TEMPERATURE=0.2 and the additional tag `"response_format": "json_object"` to create a short form answers from long form. It helps to better compare performance through an LLMs.

I FreshQA as Validation Data

Using the FreshQA test part as a validation metric may lead to suspicion about the fairness of all the results. Therefore, we split an additional 20% of the training data as a validation and retrained two models. After retraining our classifier using only the training split of our dataset – further divided into training and validation sets for hyperparameter tuning – we achieved an F1 score of 0.845 for Small E5 and 0.836 for Large E5 on English data. As shown in Table 11, FreshQA scores remain close to the original results (Table 10), though the small model performs slightly worse, likely due to the 20% data reduction. Nonetheless, the results demonstrate the generalizability of our approach.

J Error Analysis Extended

Model	Russian	English	French	German	Hebrew	Arabic	Chinese	AVG
Validation Data (FreshQA)								
BERT base cased (Devlin et al., 2019)	0.822	0.860	0.800	0.832	0.770	0.783	0.854	0.818
Deberta v3 base (He et al., 2023)	0.811	0.851	**0.841**	0.832	**0.841**	0.830	0.834	0.834
E5 Small (Wang et al., 2024)	0.809	0.839	0.818	0.830	0.801	0.815	0.794	0.815
E5 Large (Wang et al., 2024)	**0.824**	**0.872**	0.835	**0.871**	0.831	**0.835**	**0.864**	**0.848**
Test Data								
BERT base cased (Devlin et al., 2019)	0.893	0.900	0.889	0.884	0.889	0.883	0.902	0.891
Deberta v3 base (He et al., 2023)	0.836	0.842	0.845	0.841	0.832	0.825	0.831	0.836
E5 Small (Wang et al., 2024)	0.821	0.822	0.819	0.815	0.804	0.807	0.817	0.815
E5 Large (Wang et al., 2024)	**0.910**	**0.913**	**0.909**	**0.910**	**0.904**	**0.900**	**0.897**	**0.906**

Table 10: Comparison of different models on a training dataset. All models are multilingual variants. The best scores are shown in **bold**.

Model	Russian	English	French	German	Hebrew	Arabic	Chinese	AVG
Validation Data (as 20% from Train Data)								
E5 Small (Wang et al., 2024)	0.941	0.928	0.934	0.937	0.925	0.927	0.924	0.931
E5 Large (Wang et al., 2024)	0.962	0.965	0.977	0.967	0.962	0.963	0.973	0.967
Test Data (FreshQA)								
E5 Small (Wang et al., 2024)	0.796	0.845	0.803	0.755	0.766	0.762	0.748	0.783
E5 Large (Wang et al., 2024)	0.837	0.836	0.837	0.828	0.8115	0.824	0.855	0.833

Table 11: Comparison of different models on a validation split of training dataset. All models are multilingual variants.

Misclassification reason	**Example questions**
False Positives (non-evergreen, but classified as evergreen)	
Temporal phrasing mistaken for fixed historical facts	· In what year will the presidential election take place in Russia? · When will the full moon be in April?
Superlatives assumed to be static facts	· What is the biggest star in the sky? · Which tea is the healthiest? · What is the most popular social network in the world?
Biographical/life data on alive people treated as static	· In which movies has Danila Kozlovsky acted? · How many works has Stephen King written?
Geographic facts seen as immutable	· What is the length of the Amazon River? · Where is the largest zoo located?
"How-to" questions with time-sensitive/legal context	· How can maternity capital be used for building a house? · How can I contact Sberbank from abroad?
False Negatives (evergreen, but classified as non-evergreen)	
Superlatives treated as time-sensitive or trend-based	· What is the oldest currency? · The rarest element in the periodic table. · How long did the shortest war in history last?
Biological and geographical facts wrongly assumed to change frequently	· What is the area of Liechtenstein? · Which animal has the highest blood pressure? · How many species of elephants currently live on the planet?
Cultural or mythological constants treated as mutable	· Where does Ded Moroz live? · How old is Ded Moroz?
Historical events treated as recent or developing stories	· In what year was the last eruption of Mount Vesuvius? · What is the role of the French Revolution?
Recent years treated as too recent to be stable	· Who was recognized as the best actor in 2024? · Who is in first place on the Forbes list in 2024? · What is the subsistence minimum set in Russia in 2024? · What is the most popular TV series in 2023?

Table 12: Error Analysis of EG-E5 Classifier: breakdown of misclassification patterns.