FINDINGS OF EACL 2027 • TEMPORAL RAG EVALUATION

TRAVA

A Temporal RAG Validity Audit of Knowledge Cutoffs and Evidence Support

Xuan-Bach Le[1,2,*,†], Phat T. Tran-Truong[1,2,*], Minh-Hieu Le[1,2], Duy Nguyen[4], Son Ha Xuan[3]

[1]Faculty of Computer Science and Engineering, HCMUT, Vietnam [2]Vietnam National University Ho Chi Minh City, Vietnam
[3]The Business School, RMIT University, Vietnam [4]University of Economics Ho Chi Minh City, Vietnam
[*]Equal contribution. [†]Corresponding author: Xuan-Bach Le (lexuanbach@hcmut.edu.vn).

Code & cached evidence

A RAG answer is correct. But where did the answer come from? TRAVA checks three conditions:

L1 • QUESTION TIME
Did the model already know it?

4.9 → 0.2

F1 closed-book Static–Fresh gap (Qwen2.5-32B), selected vs. conservative date split. **Both uncertain.**

L2 • EVIDENCE TIME
Did the index leak the future?

11.7%

of retrieved contexts post-date asked + 3 days, yet masking them moves scores by ≤ **1.9 F1**.

L3 • EVIDENCE CONTENT
Did the shown text hold the answer?

≈ 21 F1

lost on Fresh questions with a stale pre-2025 index; Static moves < **1 F1**.

1 One real question through the three checks

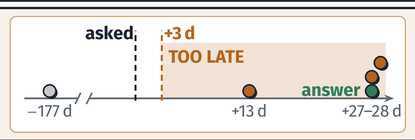
"Where will the 2026 World Cup championship match take place?"
RealTimeQA asked **2026-05-14** gold answer **New Jersey**

L1 QUESTION TIME — **Static or Fresh at boundary t_c?** The label moves when t_c moves.



FRESH at t_c = 2025

L2 EVIDENCE TIME — **Published by asked + 3 days?** Unmasked Hybrid top-5.



4/5 TOO LATE incl. the answer

L3 EVIDENCE CONTENT — **Answer in the shown text?** Ranks 3 and 5 contain "New Jersey".



2/5 SHOW IT string match

The mask removes **4 of 5** passages, including the only answer mention (**+27 days** late). After re-ranking, **2 shown passages** contain the answer:
admissibility and support are separate checks.

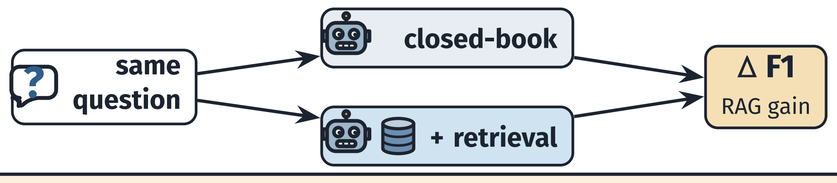
2 Paired comparisons on dated questions

878 404 Static 474 Fresh

199 weekly quizzes resampled whole

5×600 passages × chars Hybrid BM25 + BGE

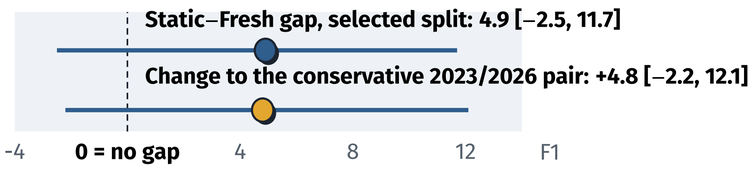
3 | 10 generators index | visibility



RealTimeQA [1], 2022–2026. Token-F1, 0–100; uncertainty resamples whole weekly quizzes. The year boundary describes **question cohorts**, not which questions a model saw in training.

3 Time checks change the interpretation

L1 • the date split moves the closed-book contrast



Qwen2.5-32B: the gap falls to **0.2 F1** under the conservative pair, but every interval crosses zero: the change is **unresolved**.

L2 • exposure is frequent; its effect is small

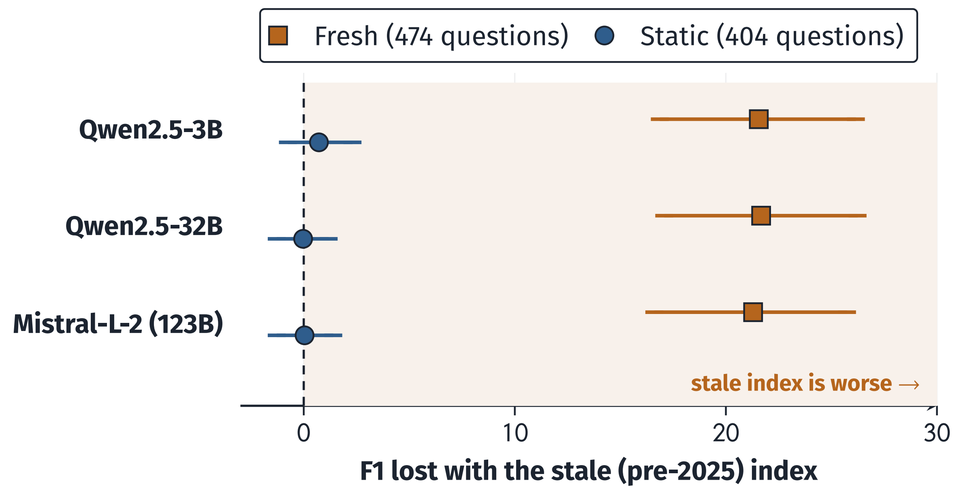
11.7% of retrieved contexts post-date asked + 3 d

34.3% of questions get at least one such passage

38% of questions change their top-5 under the mask

Yet covered generator/slice scores move by at most **1.9 quiz-weighted F1**; 18 of 20 paired intervals include zero. Equivalence within ±2 F1 holds overall across tested windows, not for every Static/Fresh subgroup.

4 L3 • A stale index costs Fresh ≈ 21 F1



Index restricted to **pre-2025 documents**: Fresh loses **21.3–21.7 F1** at 3B, 32B and 123B; Static moves by less than one point. Removing each question's linked evidence lowers **both** regimes by 18–22 F1.

Compound intervention: pool, ranking and coverage change together, so this does not isolate a causal effect of recency. Static is partly treated.

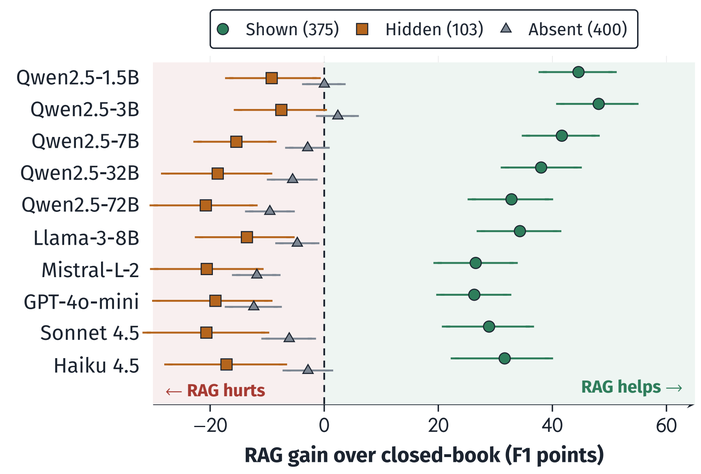
5 Retrieved is not the same as shown

Shown: in the first 600 chars

Hidden: cut off after 600 chars

Absent: in no retrieved passage

All ten generators **lose** on the **103 hidden-answer questions**: gains **45–57 F1 below** the shown group.



Points: question-pooled gains; bars: 95% quiz-bootstrap intervals.

Observational: difficulty and retrieval rank differ. Gold-conditioned support locates gains; it does not identify evidence use or give an online routing signal.

6 What a temporal RAG report should include

Boundary sensitivity: name the date split; separate cohort performance from cutoff identification.

Temporal admissibility: publication window, exposure frequency, paired masking effect.

Visible support: label the text actually shown, under the matching retrieval condition.

Intervention effects: condition-specific gains, uncertainty, multiplicity corrections.

Utility unresolved: no date-policy gain survives selection-refit inference with Holm correction.

Limits. English, Western-news source; legacy human labels use 700-character excerpts while generators see 600; exact-input human packs and Fresh tokenizer validation remain outstanding.

Code, cached evidence & reproduction: https://github.com/lexuanbach/TRAVA

Contact: lexuanbach@hcmut.edu.vn. [1] Kasai et al. *RealTime QA: What's the Answer Right Now?* NeurIPS 2023. All results follow the revised TRAVA manuscript.

Report the conditions behind the score.