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).



When a RAG answer is correct, what made it possible?

github.com/lexuanbach/TRAVA

A correct answer can have several sources

L1 L2 L3

Closed-book

The model can answer **without** the retrieved evidence.

→ **L1 question time**

Later evidence

The index holds information published **after** the question.

→ **L2 evidence time**

Shown support

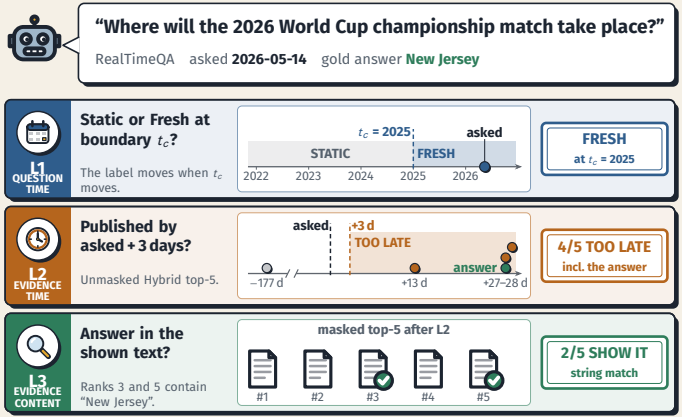
The text **actually visible** to the model contains the answer.

→ **L3 evidence content**

The same aggregate score can hide all three. **Goal:** separate date-boundary sensitivity, temporal admissibility and shown-text support *before* interpreting retrieval gains.

Three checks, illustrated by one real question

L1 L2 L3



L1 Which side of the date split?

L2 Published within the grace window?

L3 Does the *shown text* support the answer?

Re-ranking removes four later passages; two shown passages then mention the answer.

Paper Figure 1. Question dated 2026-05-14; three-day publication convention.

Keep comparisons paired within question

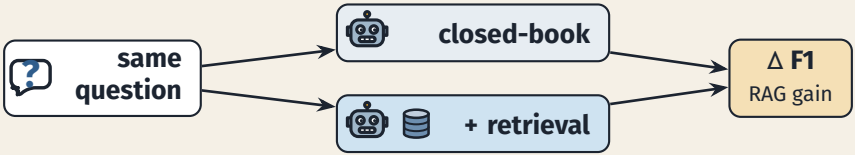
L1 L2 L3

878
RealTimeQA questions
404 Static / 474 Fresh

5×600
passages × chars
Hybrid BM25 + BGE

199
weekly quizzes
bootstrap clusters

0–100
token-F1 scale
split 2025-01-01



Closed-book, temporal-mask and index-intervention comparisons use the same questions. Generator membership is specific to each analysis.

RealTimeQA: Kasai et al., NeurIPS 2023. Paper Section 4; dates span 2022–2026.

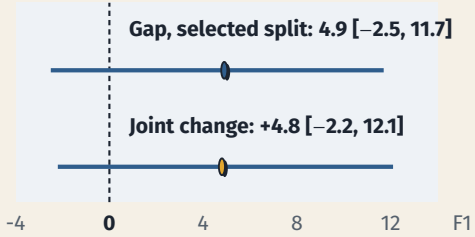
L1: the calendar split changes the contrast

L1 L2 L3

QWEN2.5-32B CLOSED-BOOK GAP

4.9 → 0.2

F1: selected split → conservative 2023/2026 pair



Both cross zero; the conservative-pair interval also includes zero.

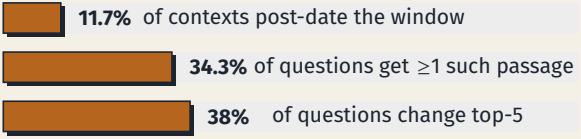
Both contrasts, and their difference, remain uncertain. Calendar cohorts do not identify a model's knowledge cutoff.

Paper Section 5.1. Corrected boundary contrasts: l1_gap_differences.json.

L2: frequent exposure, small aggregate effects

L1 L2 L3

EXPOSURE: HOW OFTEN LATER EVIDENCE APPEARS



EFFECT OF A PAIRED MASK

≤1.9 F1

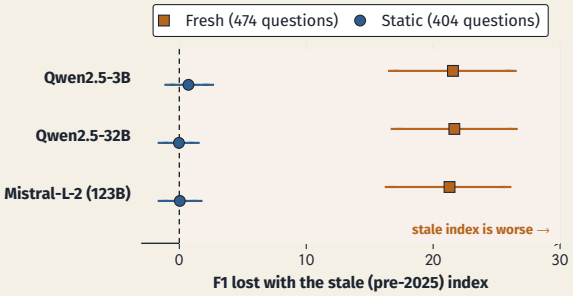
largest covered quiz-weighted change;
18 of 20 intervals include zero

Exposure is not effect size. Overall equivalence within ±2 F1 across tested windows does not establish equivalence in every subgroup.

Paper Section 5.2 and Table 1. Publication-window sensitivity: 0, 1, 3 and 7 days; five generators.

L3: a stale index costs about 21 Fresh F1

L1 L2 L3



PRE-2025 DOCUMENTS ONLY

21.3–21.7

F1 lost on Fresh

< 1

F1 moved on Static

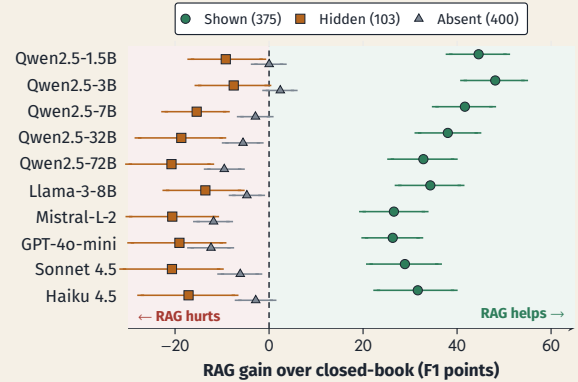
Leave-one-out lowers **both** regimes by 18–22 F1.

Pool, ranking and coverage change together: a compound intervention. Static is partly treated.

Paper Table 3. Question-pooled losses; 95% quiz-bootstrap intervals. 474 Fresh / 404 Static questions.

Retrieving an answer does not ensure it is shown

L1 L2 L3



Shown: in the first 600 chars

Hidden: cut off after 600 chars

Absent: in no retrieved passage

103 hidden-answer questions: negative gains for all ten generators, **45–57 F1 below** the shown group.

Observational: difficulty and retrieval rank differ across groups.

Paper Figure 2. Unmasked Hybrid; gold-string visibility in 600-character prefixes. Question-pooled gains, 95% quiz-bootstrap intervals. Legend counts partition 878 questions.

Support is an offline diagnostic

L1 L2 **L3**

What we measure

Does any shown passage mention or entail the gold answer?

Compare RAG gains on supported and unsupported questions.

What it does not establish

Whether the model used that evidence.

Whether a pre-answer router can observe the label.

Whether support is date-resolved.

Gains concentrate where shown text supports the answer. Gold-conditioned labels locate that pattern without identifying the mechanism.

Paper Section 5.3. Corrected labels use 600 characters and the matching retrieval condition.

Better diagnostics do not guarantee policy utility

L1 L2 L3

Conditional analysis

Some held-out gains and losses survive Holm correction.

These intervals condition on the selected policy.

Selection-refit analysis

Repeat policy selection in each bootstrap sample.

All Holm-adjusted p values are at least **0.54**.

No utility gain is resolved after selection-refit inference. The result exposes model-specific trade-offs, not a deployment improvement.

Paper Section 5.4 and Table 2. Conditional: 20 repeats; selection-refit: 5 repeats, 1,000 outer resamples.

What is validated, and what remains open?

L1 L2 L3

Available now

Public code, cached outputs, annotations and numerical results.

CPU analysis replay; no new answers or labels required.

Visible-context and masked-alignment checks.

Still limited

Human labels: 700-character excerpts; generator input: 600.

Exact-input human packs remain **uncollected**.

Fresh tokenizer validation and independent temporal/domain generalization remain open.

Paper Limitations and reproducibility appendix. MIT covers code; third-party data have separate terms.

Report the conditions behind the score

L1 L2 L3

Question time: show boundary sensitivity.

Evidence time: separate exposure from effect.

Evidence content: label what was shown.

Pair these checks with index interventions and uncertainty.

Code & cached evidence

https://github.com/lexuanbach/TRAVA

lexuanbach@hcmut.edu.vn

Stale-index losses are large; shown support locates retrieval gains. Neither result alone identifies the model's evidence-use mechanism.

Backup: human-validation counting units

L1 L2 L3

Study	Counting units and interpretation
Task B	900 passages; 300 question–context instances; 254 unique questions.
Masked support	630 passages; 210 questions = 200 sampled + 10 consistency cases.
Population estimate	Excludes the ten cases without design weights; interactions on all 210 are sample-level summaries.
Exact-input packs	900 + 630 rows prepared; no new human labels collected.

At the token-F1 threshold of 0.5, the conditional miss rate among human-correct outputs is 24.0% [9.6, 38.3]. This is not the proportion of all outputs missed.

Paper Sections 4 and 5.3; human-validation appendix. Metric audit: 395 items, adjudicated labels and common resampling.

Backup: scope and evidence pointers

L1 L2 L3

Within-source replication: 800 unused questions; 172 of 182 quiz weeks overlap the original sample. Not independent temporal generalization.

Context budget: the corrected ten-passage prefix arm retains the original five passages and changes support membership. Fixed-partition differences remain unresolved.

Mistral format: cached runs use a fallback chat template. Native-template sensitivity lowers its Static support interaction; magnitudes are configuration-specific.

Evidence in the artifact

`docs/RESULTS.md`

`results/analysis/stale_placebo.json`

`results/analysis/hidden_answer_contrast.json`

`results/analysis/r4_e5_outer.json`

Dataset: Kasai et al. *RealTime QA: What's the Answer Right Now?* NeurIPS 2023.