TRAJBL
EVIDENCE THAT SCALES

Trajbl Public Evidence Brief v0.1

Expanded metrics edition - public evidence summary for technical review, partnership screening, and NDA discussions.

64-88%
reported input context reduction across Phase29 slices

95%+
Croatian quality retention vs full-context RAG

0.7895
Yes/No Bridge F1, rank #1 in public leaderboard

100%
Trajbl extraction success rate on HR Provence comparison

What Trajbl is.

Trajbl is a proprietary, question-aware evidence-packing layer for RAG/LLM systems. It sits after retrieval and before generation, selecting whole, auditable sentences as compact evidence rather than pruning individual words or sub-word tokens.

Positioning.

The public evidence supports a focused claim: Trajbl can sharply reduce input context while retaining strong QA quality on the reported Croatian and English slices, and it shows clear wins in several specialized categories: binary Yes/No, relational multi-hop queries, Croatian Exact Match, extraction robustness, CPU-only deployment, and citation auditability.

Tone boundary.

This brief does not claim SOTA, peer review, independent validation, full reproducibility, or overall superiority against every compressor. It presents the strongest public results with their scope and caveats.

Prepared for: technical reviewers, AI infrastructure teams, RAG platform teams, and potential partners

Contact: Amir Šerbić - amir.serbic@yahoo.com

Public repo: https://github.com/Zivotu/Trajbl_Public_Evidence

Public landing: https://neurobiz.me | Controlled demo: https://neurobiz.me/demo/

1. Where Trajbl sits in the RAG/LLM pipeline

Trajbl is not the final answer generator. It is the compact evidence layer placed between retrieval and the LLM. Its job is to preserve the strongest answer-supporting sentences while shrinking noisy or oversized retrieved context.

Positioning: post-retrieval, pre-generation. Trajbl compresses/organizes evidence before the prompt is sent to the LLM.

Core design choices

Design choice	Practical meaning	Why it matters
Whole-sentence extraction	Keeps complete, contiguous sentences instead of token fragments.	Better readability, safer citation, lower risk of deleting negations or boundary conditions.
Question-aware selection	Ranks evidence by relation to the user query and context structure.	Reduces context dilution and focuses the generator on relevant evidence.
Source-balance strategy	Distributes evidence across documents when cross-doc reasoning is needed.	Prevents one dominant source from crowding out minority evidence.
Deterministic / CPU-only runtime	No GPU model inference is required at the compression step.	Predictable behavior, simpler deployment, lower overhead.
Auditability	Packed sentences remain human-readable and linkable to source text.	Useful for regulated or evidence-sensitive domains.

Distinction from token pruning

Token-level pruning systems can be efficient, but the public repo argues that deleting words or sub-word tokens can produce broken evidence fragments. Trajbl's positioning is different: it compresses by selecting complete evidence units, not by shaving syntax inside the sentence.

2. Phase29 core results - context reduction with scoped quality retention

The reported Phase29 slices measure word count as a proxy for context budget. The strongest Croatian claim is not merely 'smaller context', but smaller context while retaining 95%+ of full-context RAG quality on the tested Croatian slices.

63.6%
HR single-doc context reduction

82.5%
HR cross-doc context reduction

87.6%
EN LongBench-E context reduction

97.4%
HR cross-doc quality retention

Slice	n	Full RAG	Trajbl	Input words	Reduction	Retention / note
HR single-doc QA	109	0.661	0.628	3183 -> 1160	63.6%	95.0% retention
HR cross-doc QA	16	0.625	0.609	7456 -> 1308	82.5%	97.4% retention
EN LongBench-E	100	0.587	0.400	10593 -> 1313	87.6%	Early English signal; SQuAD F1

Interpretation

The Croatian Phase29 results are the cleanest headline: Trajbl compresses substantially while staying close to full-context RAG quality. The English LongBench-E result is valuable as an early English-language validation signal, but it should not be sold as a broad English proof across all domains.

! Caveat: Croatian quality scores rely on an LLM-judge rubric and the cross-doc slice is small (n=16). These are evidence signals, not peer-reviewed conclusions.

3. Reported wins against Microsoft Research LLMLingua-2

The public evidence supports a clear scoped claim: Trajbl outperforms all reported LLMLingua-2 baselines/slices in the Phase29 public tables. This includes question-aware and question-blind LLMLingua-2 where both are reported.

Benchmark	Trajbl	LLMLingua-2 baseline	Delta	Context budget note
HR single-doc vs Q-aware	0.628	0.438	**+0.190**	Trajbl: 1160 words; LLMLingua-2: 1331
HR single-doc vs Q-blind	0.628	0.420	**+0.208**	Trajbl: 1160 words; LLMLingua-2: 1290
HR cross-doc vs Q-aware	0.609	0.313	**+0.297**	Trajbl: 1308 words; LLMLingua-2: 1350
EN LongBench-E vs Q-aware	0.400	0.233	**+0.167**	Trajbl: 1313 words; LLMLingua-2: 1405

Why this matters

LLMLingua-2 is a well-known Microsoft Research context compression baseline. Beating it in every publicly reported Phase29 comparison is one of Trajbl's strongest external-positioning points. The clean wording is: Trajbl outperformed LLMLingua-2 on the reported Croatian QA and English LongBench-E slices under the reported compressed-budget setup.

Boundary: the public repo does not prove global superiority over every LLMLingua-family configuration, every LongLLMLingua setup, or every task type. Do not oversell it as 'beats all Lingua tools everywhere'. The supported claim is strong enough without that mistake.

4. Compression leaderboard - where Trajbl ranks first in the public table

The updated public repository includes a compact leaderboard comparing Trajbl with Provence, LLMLingua-2, and naive truncation across operational and task-specific criteria. The most important point: Trajbl is not only a compressor; it is strong where evidence structure matters.

Dimension	Trajbl	Naver Provence	LLMLingua-2	Naive truncation
Context savings	**64-88%**	~65-80%	~60-80%	~60%
Yes/No binary QA	**0.7895 F1 - Rank #1**	0.1579 F1	<0.15 F1	<0.10 F1
Relational multi-hop QA	**0.4450 F1 - Rank #1**	0.3310 F1	<0.20 F1	<0.15 F1
Exact Match (EM)	**0.10 - Rank #1**	0.06	<0.05	<0.05
Extraction robustness	**100% - Rank #1**	87.5%	100%	100%
Compute overhead	**CPU-only; zero GPU**	GPU recommended	GPU required	CPU-only
Auditability	**Contiguous sentences**	Incomplete fragments	Broken fragments	Arbitrary truncation

Best public-use sentence

◎ Trajbl ranks first in the public compressed-setup leaderboard on Yes/No binary QA, relational multi-hop QA, Croatian Exact Match, extraction robustness, CPU-only operation, and sentence-level auditability, while still preserving strong context-budget reduction.

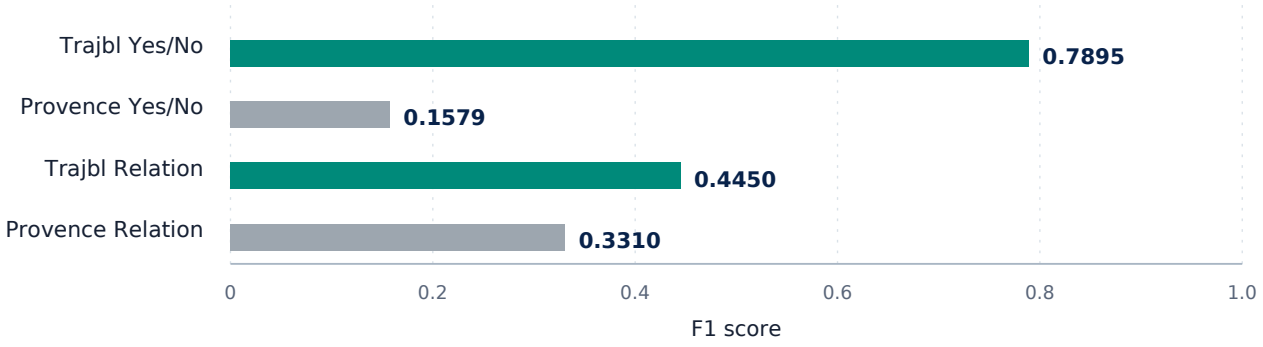
✓ Boundary: this leaderboard is public evidence, not an independent benchmark paper. Keep the table, but keep the caveats near it.

5. Provence comparison - wins, losses, and the honest story

This is where the brief must be careful. Provence retains a semantic F1 advantage on generic QA slices. That does not erase Trajbl's real wins: robustness, Exact Match on Croatian, Yes/No bridge performance, relational bridge performance, deterministic CPU-only execution, and citation auditability.

Metric / slice	Trajbl	Provence	Interpretation
English generic F1	0.4532	0.4914	Provence leads generic semantic F1
Croatian generic F1	0.4047	0.4388	Provence leads generic semantic F1
Croatian EM	0.10	0.06	Trajbl leads exact-match precision
Extraction success	100%	87.5%	Trajbl had zero empty outputs; Provence failed 6/48
Yes/No Bridge, n=19	0.7895 F1	0.1579 F1	Trajbl +0.6316 on specialized binary tasks
Relation Bridge, n=21	0.4450 F1	0.3310 F1	Trajbl +0.1140 on relational multi-hop tasks

Visual check - specialized wins



The strongest honest positioning: Provence is stronger on broad semantic relevance; Trajbl is more attractive where deterministic, CPU-only, sentence-level, auditable evidence packing and specialized structural query behavior matter.

6. Advanced modules and newer evidence signals

The public repo includes two additional technical signals beyond Phase29. These should be presented as promising engineering evidence, not as broad final proof.

Module / evidence	Result	Why it matters	Caveat
Confidence Gate routing - Phase56	0.7225 vs 0.7041; +0.0184; 2W/0L/107T	Selective routing can improve quality without observed regressions on the tested 109-question slice.	Needs validation across more domains and query distributions.
Value-Point Repair - Phases60-66	0.7979 vs 0.7606; +0.0372; 6W/2L/39T	Promising for numeric, value-heavy and marker questions where exact values matter.	Research/protected candidate; non-default at runtime because gains are slice-specific.

Practical fit

Good fit	Why
Large-context RAG systems	Reduces context size before the expensive LLM step.
Evidence-sensitive domains	Whole sentences support audit trails and citation review.
CPU-only or low-resource deployments	Selection step does not require GPU/VRAM.
Cross-document QA	Source-balance strategy is designed to reduce source crowding.
Binary, relational, numeric queries	Public slices show specialized bridge / repair gains.

7. Evidence boundaries, links, and contact

This brief is useful precisely because it separates strong public claims from claims that are not yet supported publicly. That boundary protects credibility.

Boundary	Plain-language meaning
Not peer-reviewed	The results have not been independently peer-reviewed by an academic or industry committee.
Source code proprietary	Code, scoring boundaries, routing files, cue lists, coefficients, and internal prompts are not public.
Not a full reproducibility package	The repository contains aggregate/sanitized evidence, not raw datasets, reference answers, or full transcripts.
Provence caveat	Provence retains a semantic F1 advantage on generic QA slices.
LLMLingua scope	The strong public claim is against reported LLMLingua-2 baselines/slices, not every possible Lingua-family configuration.
Demo caveat	The controlled demo demonstrates behavior; it is not a benchmark harness.

Public links

Public evidence repository: https://github.com/Zivotu/Trajbl_Public_Evidence
Public landing page: https://neurobiz.me
Controlled demo: https://neurobiz.me/demo/

Contact / NDA note

For technical validation, private evaluation, source review, or partnership discussions, contact Amir Šerbić at amir.serbic@yahoo.com. Detailed code, parameters, and private benchmark materials can be discussed only through a formal NDA or partner-review process.

Source basis for this brief

Metrics are drawn from the public repository files: README.md, BENCHMARK_SUMMARY.md, CLAIM_LEDGER.md, LIMITATIONS.md, METHOD_OVERVIEW_NO_CODE.md, and RESULTS_TABLES/*.csv, checked against the public GitHub repository on 8 June 2026.