GEO-PTBR: A Brazilian-Portuguese Replication of Generative Engine Optimization and the Engine-Dependence of Its Effects

Elio Suraci Picchiotti
AEO BR, Caracol Media — Brazil
elio.picchiotti@aeobr.com.br
ORCID: 0009-0006-3058-4096

August 13, 2026

Abstract

Generative engines answer queries by synthesizing and citing a few retrieved sources, displacing the link list and with it the mechanics by which content creators earn visibility. Aggarwal et al. [2024] introduced *Generative Engine Optimization* (GEO): nine content-side transformations that raise how prominently a source is cited, with gains up to 40% in English. This paper reports GEO-PTBR, the first replication in Brazilian Portuguese, and turns cross-engine agreement from a robustness check into the measurement itself. Holding retrieval fixed, we apply each technique to one of five already-retrieved PT-BR sources over 525 queries in three sectors, and run the *same* queries and the *same* transformed sources against three economy-tier engines from three model families, so that which engine answers is the only variable that changes.

The effect of a GEO technique depends on who answers. The three engines agree on the direction of the effect for only four of nine techniques (44.4%), and all four are techniques that *hurt*: what transfers is which optimizations backfire, not which ones work. Under Holm correction over the primary family of 27 tests, Fluency Optimization inverts outright: $+7.9\%$ on one engine, -6.9% on another, both arms surviving. The most reactive engine responds 5–$17\times$ more strongly than the least. Against the original English results, five of nine directions replicate but the level does not: its top techniques gain $+27$ to $+41\%$; our largest positive effect is $+2.6\%$. Rank correlations of technique ordering, under the primary (median) estimator, do not reach significance in an exact permutation test, so we claim no transfer of ordering. Five of nine techniques induce citation of sources that cannot answer the query — a safety failure, not an optimization. Sources are model-generated texts, not live web pages; dataset, traces and code are released in full.

1 Introduction

Large language models have enabled a new class of search system — *generative engines* (GEs) such as Google's AI Overviews, Bing Copilot, and Perplexity.ai — that answer a user's query directly, in prose, with inline citations to a handful of retrieved sources, rather than returning a ranked list of links. This shift is convenient for users but removes the traditional mechanism (organic ranking) by which content creators earned visibility and traffic: a source's owner has little insight into, or control over, when and how their content is surfaced inside a generated answer. Aggarwal et al. [2024] formalized this problem as *Generative Engine Optimization* (GEO) and proposed nine content-side transformations — applied via prompting to a single retrieved source, without touching the retrieval step itself — that can increase how much of a generated answer cites that source, reporting visibility gains of up to 40% on their English-language GEO-bench benchmark.

Brazil is not a marginal case for this shift. In the Reuters Institute's 2026 Digital News Report, weekly use of AI chatbots for news reaches 13% of Brazilian respondents against a 10% global average [Egan et al., 2026], placing the country above the international trend rather than behind it — and the language in which those answers are generated is Portuguese.

In Brazil, this is also no longer only a content-strategy question. On April 23, 2026, CADE — the Brazilian competition authority — unanimously opened a formal administrative proceeding against Google over the use of journalistic content in AI-generated summaries, on evidence of exploitative abuse of a dominant position [Conselho Administrativo de Defesa Econômica (CADE), 2026]. What makes the case relevant to a measurement paper is the specific mechanism the tribunal identified: the existing opt-out controls were characterized as a *false choice*, because a publisher who refuses to appear in AI summaries cannot do so without losing search visibility it depends on commercially. The quantity at issue in that proceeding — how much of a generated answer a given source gets, and what the publisher can do about it — is the quantity this paper measures, on Brazilian-Portuguese content, in the market where the proceeding is running. We take no position on the antitrust question, and nothing in our design was chosen to inform it; we note the proceeding only to establish that content-side visibility in generative answers is a live, contested variable in this market rather than a hypothetical one.

Two gaps motivate this work. First, GEO's original evaluation is entirely in English, over English-web sources and English-native query distributions (MS MARCO, Natural Questions, ELI5, and similar); whether the same techniques help, hurt, or do nothing in a different language and market — Brazilian Portuguese, in our case — is an open empirical question, not something safe to assume by analogy. Second, subsequent work has complicated the picture: Puerto et al. [2025] evaluated a wider set of "conversational SEO" strategies across six domains and roughly 1,900 queries and found that the large majority of method–domain combinations produced no significant, or actively harmful, effect once techniques are evaluated under more realistic multi-actor and competitive conditions — a different experimental regime from GEO's single-source-optimized setting (see §2 for how this project positions itself relative to both).

GEO-PTBR reproduces, in laboratory conditions, the final stage of a generative engine's pipeline: given a fixed set of five already-retrieved PT-BR sources for a query, which source(s) does the model cite, and how does that change when one source is optimized with one of the nine GEO techniques. This isolates the causal effect of content-side optimization on the citation/summarization step, holding retrieval fixed — it does *not* measure whether an optimized page would have been retrieved (or crawled, or indexed) in the first place. Two boundaries of the object under study should be stated before any result is read, not discovered in the Limitations. The sources are model-generated PT-BR texts, labeled as such and released in full: our results are about *optimized text*, not about live Brazilian web pages with their HTML, markup, boilerplate, navigation and domain authority, none of which this design contains. And the five sources per query are a fixed context, so "visibility" here means share of a generated answer, never traffic. See §6 for the full scope of what this design does and does not measure.

Contributions.

1. A public PT-BR benchmark of 525 real-user-register queries across three sectors — health, legal, and real estate — with 5 sources each (150–400 words, PT-BR), including a 25-query pitfall control set (queries the given sources cannot answer) used to detect measurement hallucination, released with per-technique results for every engine.
2. A PT-BR adaptation of all nine GEO techniques, implemented as fixed, versioned transformation prompts (§3), applied to a per-query target source whose position is deterministic but held constant across all nine techniques, following the original paper's protocol.
3. A *three-engine* evaluation in which the full benchmark — the same queries, and the same already-transformed source texts — is run against `gemini-3.5-flash-lite`, `claude-haiku-4-5`, and `gpt-5.6-luna`, three engines from three different model families, with 3 repetitions per cell. Because the transformation step is performed once and reused, the only variable that changes across the three runs is which engine produces the answer, which isolates engine-side citation behavior from transformation quality.
4. Evidence that the effect of a GEO technique is **engine-dependent**: techniques that measurably help on one engine measurably hurt on another, over the same queries, with bootstrap confidence intervals excluding zero on both sides. This is a finding about the transferability of the GEO

protocol itself, not a robustness check on our own numbers, and it is reported with an explicit direction-agreement rate over all nine techniques (§5.2).

5. A direct, direction-of-effect comparison against the original GEO paper's published tables, using rank correlation rather than absolute values (the two studies use different engines and different normalization, so absolute visibility scores are not comparable across them; see §3).
6. A public release of the full dataset — queries, sources, per-technique results for every engine, *and the raw execution traces* — at `https://huggingface.co/datasets/epicchi2103/geo-ptbr`, under CC BY 4.0. Releasing the traces is deliberate: this study holds that a result exists only if it came from a recorded execution, and every number in this paper can be recomputed from those traces with the released code (`https://github.com/epicchiotti2103/geo-ptbr`, MIT).

2 Related Work

Generative Engine Optimization. Aggarwal et al. [2024] introduced GEO and GEO-bench, a 10,000-query benchmark spanning nine source query datasets (MS MARCO, ORCAS-1, Natural Questions, AllSouls, LIMA, Davinci-Debate, Perplexity.ai Discover, ELI5, and GPT-4-generated queries) and 25 topical domains. For each query, five retrieved sources are placed in a fixed context window and a generative engine (primarily `gpt3.5-turbo`, prompted — not fine-tuned) produces a cited answer; one randomly chosen source is then optimized with each of nine techniques in turn, and visibility is measured with two objective metrics (word count and position-adjusted word count, §3) plus a subjective, G-Eval-based impression score. Their headline finding is that three techniques — Cite Sources, Quotation Addition, and Statistics Addition — each add 30–40% position-adjusted visibility on average, that purely stylistic changes (fluency, simplification) also help by 15–30%, that classic SEO tactics (keyword stuffing) hurt rather than help, and that gains are far larger for sources ranked poorly in the underlying search index (rank 5) than for already top-ranked sources (rank 1) — a result they read as evidence that GEO has a "democratizing" effect for small or poorly-ranked publishers. They additionally validate a subset of these findings against a deployed commercial engine, Perplexity.ai. Full details of every table and finding we treat as our target of comparison are recorded in `paper/KEY_FACTS.md` in this project's repository, along with several internal inconsistencies in the original paper's own tables (e.g., a caption percentage that does not exactly match the table it captions) that we note but do not attempt to resolve.

C-SEO Bench. Puerto et al. [2025] extend this line of inquiry with C-SEO Bench, evaluating conversational-SEO-style techniques across two tasks, six domains, roughly 1,900 queries, and over 16,000 documents. Critically, their design allows for *competitive*, multi-actor adoption — multiple sources in the same context can be optimized simultaneously — closer to what would happen if GEO techniques were adopted at scale. Under that regime, they find that only 3 of 54 method–domain combinations produce a significant positive effect, and none does in question-answering tasks, suggesting several widely cited "GEO hacks" may not survive realistic, adversarial adoption. GEO-PTBR follows the original GEO paper's *single*-source-optimization protocol, not C-SEO Bench's competitive one (see §3): our results should therefore be read as a same-paradigm replication of Aggarwal et al. [2024] in PT-BR, not as a competitive-adoption stress test in the style of Puerto et al. [2025]. We take two lessons from C-SEO Bench regardless of protocol choice: (i) that null and negative results are common and worth reporting honestly rather than only highlighting positive findings, which motivated our explicit "what did NOT work" finding (§5); and (ii) that method effectiveness is domain- and context-dependent, which motivated our sector-level breakdown (health / legal / real estate) rather than a single pooled result.

Learned and adaptive alternatives to fixed techniques. A recent line of work abandons the fixed catalogue of hand-designed techniques altogether. Wu et al. [2026] introduce AutoGEO, which prompts

frontier LLMs to explain a generative engine's preferences, extracts preference rules from those explanations, and uses them both as context engineering for a prompt-based system and as rule-based rewards to train a smaller model; they report that the learned rules capture preferences that are specific to different domains. Yuan et al. [2026] propose AgenticGEO, which motivates its design by arguing that existing methods rest on static heuristics or preference-rule distillation prone to overfitting, and therefore "cannot flexibly adapt to diverse content or the changing behaviors of generative engines"; it evolves compositional strategies instead and reports transferability across two engines.

Both proceed from a premise that our results supply direct evidence for. If a fixed catalogue of content transformations carried a stable effect from one engine to the next, the case for automatically learning or evolving engine-specific strategies would be considerably weaker. We measure that premise rather than assume it: the nine techniques of Aggarwal et al. [2024], applied unchanged to identical sources and queries, agree in direction across three engines for only four of nine, and the mildest technique in the catalogue inverts sign between two engines with both arms surviving multiplicity correction (§5.2). Our study and this line of work are complementary — they propose adaptive methods, we quantify the non-transferability that motivates them — and we make no claim about the cross-engine behavior of their methods, which we did not evaluate.

Critical surveys of the field. Martinez [2026] reviews 45 GEO studies published between 2023 and 2026 and concludes that GEO is not a single ranking task but a stochastic, partially observable pipeline, and that "generic heuristics transfer poorly" across the reviewed corpus. Its strongest negative claim is precisely scoped, and we reproduce that scope rather than round it off: no reviewed technique shows a stable, longitudinal, cross-platform causal effect *on organic discoverability or downstream behavior*. The survey does *not* deny that already-retrieved content can causally alter its own citation — it grants exactly that, and it is exactly what we measure.

Our contribution is the direct experimental counterpart, one level below that claim. The survey finds cross-platform stability undemonstrated in the literature for discoverability; we measure the instability itself at the citation step, where the causal effect is granted, under controlled conditions: identical queries, identical optimized sources, three engines, and the same technique inverting sign between two engines with both arms surviving multiplicity correction (§5.2). We also extend it along an axis the reviewed corpus does not cover, since that corpus is overwhelmingly English-language.

Native PT-BR evaluation resources. Our decision to build a PT-BR corpus rather than translate an English one follows established practice in Brazilian-Portuguese IR and NLP evaluation. Bueno et al. [2024] introduce Quati, a Brazilian-Portuguese retrieval dataset built from queries written by native speakers over documents from high-traffic Brazilian sites, motivated precisely by the shortage of non-translated PT-BR IR resources; before it, PT-BR retrieval work leaned on machine-translated collections such as mMARCO. Stekel [2026] take the constraint further: MTEB-BR admits only data created or found in Portuguese and *excludes translations by construction*, across 22 tasks and 93 evaluated models. Two of their findings bear directly on our design. First, a native benchmark measures something the multilingual leaderboards do not — a model's global rank predicts its Portuguese rank only moderately, and one model ranks 3rd globally and 49th there. Second, that is the same failure mode we report one level up the pipeline: performance measured on one distribution does not carry to another, whether the shift is of language (their case) or of answering engine (ours).

Two caveats on how we use this precedent. Neither resource is a GEO benchmark — both evaluate retrieval and embedding quality, not content-side optimization of an already-retrieved source — so they license our corpus *construction* choice, not our metrics or protocol. And unlike theirs, our sources are model-generated rather than collected from the live Brazilian web (§6): our corpus is native PT-BR in register and topic, but it is not a sample of real Brazilian pages, and we do not claim it is.

3 Methodology

GEO-PTBR reproduces the GEO protocol of Aggarwal et al. [2024] on a PT-BR dataset, adapting only what is required to make the study reproducible and auditable end-to-end; every methodological decision not fully specified by the original paper is registered here and versioned in code (a change to any of these decisions increments `versao_experimento` in `config/study.yaml`).

3.1 Protocol

For each query in the dataset:

1. **Scenario.** One PT-BR query paired with 5 PT-BR sources (150–400 words each) that answer it, numbered 1–5.
2. **Baseline.** A fixed "AI search engine" prompt places all 5 sources in context and asks the engine to answer concisely, citing the source of every claim inline as `[1]`–`[5]` (multiple citations per sentence are allowed, e.g. `[1][3]`), and to state explicitly, without citing anything, if the given sources do not answer the question. Visibility of each source in the baseline answer is measured with Imp_{wc} and Imp_{pwc} (§3.2).
3. **Intervention.** One of the 9 GEO techniques (§3.3) is applied to a single, per-query *target source* (§3.4); the same call is re-run with the transformed source substituted in place of the original, and visibility is re-measured.
4. **Aggregation.** Results are aggregated by technique, by sector, and by target-source position (§3.5).

Each (query, technique) cell is repeated `3` times at temperature `0.7` (engine stochasticity; a single sample is not treated as a result) — a deliberate deviation from the original paper's 5 seeds, made for cost reasons under this study's US$40 ceiling (§4); we flag this explicitly as a limitation (§6) rather than presenting it as equivalent statistical power. The transformation itself (source → optimized source) is applied once, at temperature `0`, since it is a deterministic treatment step, not part of the stochastic measurement.

3.2 Visibility metrics

We reuse, without modification, the two objective visibility metrics defined by Aggarwal et al. [2024] (their §2.2.1). Given a generated answer r split into sentences $s \in S_r$, and $S_{c_i} \subseteq S_r$ the subset of sentences citing source c_i, with $|s|$ the word count of sentence s (citation markers excluded, and a sentence cited by multiple sources splitting its word count equally among them):

$$\text{Imp}_{wc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s|}{\sum_{s \in S_r} |s|} \tag{1}$$

$$\text{Imp}_{pwc}(c_i, r) = \frac{\sum_{s \in S_{c_i}} |s| \cdot e^{-\text{pos}(s)/|S_r|}}{\sum_{s \in S_r} |s|} \tag{2}$$

Relative improvement of a technique for source s_i is computed as in the original paper's Eq. 4:

$$\text{Improvement}_{si} = \frac{\text{Imp}(s_i, r') - \text{Imp}(s_i, r)}{\text{Imp}(s_i, r)} \times 100 \tag{3}$$

where r is the baseline answer and r' the answer generated after applying a GEO technique to s_i.

Versioned interpretation of $\text{pos}(s)$. The original paper's Eq. 3 (our Eq. 2) uses a term $\text{pos}(s)$ that it motivates qualitatively (*"sentences that appear first in the response are more likely to be read... the exponent term... gives higher weightage to such citations"*, citing power-law click-through-rate results) but never formalizes mathematically in the visible text (documented in `paper/KEY_FACTS.md` §1.2). We fix this ambiguity, once and for all analyses in this paper, as $\text{pos}(s) = \text{idx}(s)$, the 0-indexed position

of sentence s among the response's sentences — i.e. the position weight is $e^{-\mathrm{idx}(s)/|S_r|}$, where $|S_r|$ is the number of sentences in the response (written n in the implementation) — matching both the qualitative description and the reference implementation's exponential-decay-by-order behavior (implementation: `src/metrics.py`, `METRICS_VERSION = "v1"`). This is a documented interpretive choice, not a claim about what the original authors intended beyond what their text specifies; a different reasonable interpretation could shift Imp_{pwc} values, which is exactly why it is versioned.

3.3 The nine GEO techniques

We implement all nine techniques from Aggarwal et al. [2024] as PT-BR transformation prompts, applied by an LLM to one source's text at a time. Table 1 lists them; see `src/transform.py` for the exact, versioned prompt text (`PROMPTS_VERSION = "v1"`) — a change to any instruction increments this version and, transitively, the overall experiment version.

Technique	Transformation
Authoritative	More persuasive, authoritative tone; facts unchanged.
Statistics Addition	Replace qualitative claims with plausible quantitative statistics.
Keyword Stuffing	Repeat query keywords throughout, classic-SEO style.
Cite Sources	Attribute claims to plausible, trustworthy sources.
Quotation Addition	Add direct quotations attributed to plausible experts/sources.
Easy-to-Understand	Simplify language: shorter sentences, everyday vocabulary.
Fluency Optimization	Improve flow and readability without changing content.
Unique Words	Prefer rare/distinctive words over generic terms.
Technical Terms	Add domain-specific technical terminology.

Table 1: The nine GEO techniques, as adapted for PT-BR transformation prompts. One-line summaries; exact instruction text in `src/transform.py`.

Following Aggarwal et al. [2024], each technique is applied *on the same target source*, so that its effect can be isolated from confounds like source content or position: "the source is chosen randomly but remains constant for a particular query across all GEO methods."

Transformation model and self-preference. Transformations are applied by the same model family as the primary evaluation engine (`gemini-3.5-flash-lite`), matching the original paper's own choice to use `gpt3.5-turbo` for both roles. This introduces a self-preference risk — the engine could systematically favor content it (or a closely related model) generated — which we discuss and partially, but not fully, mitigate in §6.

3.4 Target-source selection

For each query, the target source's position is set deterministically as `random.Random(query_id).randint(1, 5)` (implementation: `src/run_case.py`) — a pseudo-random draw seeded by the query's own ID, so it is fixed and reproducible across runs, yet effectively arbitrary with respect to source content, matching the original paper's "chosen randomly but remains constant" property while giving this study exact reproducibility (a genuinely non-deterministic draw could not be re-derived from the dataset alone). Positions are rotated across the dataset by construction, allowing the position-level breakdown in §5.4 to be computed without confounding technique effects with a fixed position.

3.5 Statistical treatment

For each (technique, metric, subset) cell, we first aggregate the 3 repetitions *within* each query (mean), then compute a paired bootstrap *across* queries (10,000 resamples with replacement, fixed seed 42, 95% percentile confidence intervals) — aggregating by query before resampling avoids pseudo-replication

from the repeated-measures structure. Each bootstrap resample yields, from the same resampled set of queries, the baseline mean, the technique mean, the mean relative improvement, and the median relative improvement (Eq. 3), so the four statistics stay paired rather than being computed from independently resampled sets. We report both the mean and the median relative improvement: the pilot phase (§4) showed the mean of Eq. 3 to be unstable when the baseline visibility of a query is small. The median is the more robust summary statistic and is treated as primary; the mean is reported alongside it for direct comparability with the original paper's Eq. 4-based reporting.

Zero-baseline queries and the primary analysis set. Eq. 3 is undefined only when the baseline is exactly zero *and* post-technique visibility is positive. We report that case as "new" (a source not cited at all before the technique) and never impute a numeric value for it. The other zero-baseline case — the target source cited neither in the baseline *nor* under the technique — is not undefined but *degenerate*: Eq. 3 evaluates to exactly 0%, so the query enters the distribution as "no change" when what it actually records is "never cited at all." These encode two different questions — *was* the source cited, and *by how much* did its citation change — and pooling them is not harmless. Because such queries all contribute the single value 0%, they pull the median toward exactly zero, and they do so at engine-dependent rates, which makes a real negative effect on one engine readable as absence of effect.

We therefore report every Eq.-3-based table over two explicitly labelled query sets:

- **`baseline_pos` (primary)** — queries whose target source has strictly positive baseline visibility for the metric in question. Eq. 3 is always defined on this set, so no query is discarded by indeterminacy and none contributes a degenerate zero.
- **full set (sensitivity)** — all non-pitfall queries, reported alongside the primary set rather than instead of it, so that a reader can see how much of any effect is carried by never-cited queries.

Where the two disagree, the disagreement is itself informative: it is about *whether the source is cited at all*, not about the magnitude of the effect, and both rows should be read together.

A second, independent property motivates the same set for the cross-engine comparison of §5.2. On the full set, each engine discards a *different* subset of queries through the "new" case above, so a comparison that was paired by query becomes unpaired again at the statistics stage; on `baseline_pos` the effective n is identical across engines by construction. Every table reports the effective n per engine so this is visible rather than assumed. This decision is implemented once, in `src/aggregate.py` and `src/compare_engines.py`, which share the same definition. The share of queries this excludes is strongly engine-dependent, which is precisely why the distinction cannot be left implicit: 3 of 500 on `gemini-3.5-flash-lite`, against 98 of 500 on the intersection used for the cross-engine comparison. The consequence is concrete. On `claude-haiku-4-5`, Technical Terms reads as -0.8% with a confidence interval crossing zero when degenerate queries are pooled in — that is, as no effect — and as -4.4% with the interval excluding zero once they are separated. The pooled figure is not a different estimate of the same quantity; it is a different quantity, and reporting it as the effect of the technique would have turned a distinguishable negative arm into an apparent absence of effect.

Fair-comparison rule. The original paper reports visibility on an arbitrary, per-study normalized scale (its own baseline is ≈19.3–24.7 depending on the table; ours is on a $[0, 1]$ share of total answer word count, by construction of Eqs. 1–2). **Absolute values are never comparable across the two studies.** Every comparison against the original paper in this work (§5.5) uses only sign (did visibility go up or down) and Spearman rank correlation across techniques — never a ratio or difference of absolute values between the two studies.

4 Experimental Setup

Dataset. 500 PT-BR queries (design target; see §5 for the number actually executed at the time of writing) across three sectors — health (`saude`), legal (`juridico`), and real estate (`imobiliario`) — roughly 167 queries each, plus 25 "pitfall" queries (≈5% of the total) that the five given sources

cannot answer, by design, and are used as a measurement-hallucination control: if a pitfall query shows non-trivial citation visibility for any source, that flags a bug in the measurement pipeline rather than a real finding. Health is treated as a YMYL ("Your-Money-Your-Life") sector; reporting sector-level behavior differences, particularly for health, is a target finding of this study. Queries are phrased in the register of a user asking a generative AI assistant a real question, not as SEO-style keyword strings (e.g., "*Posso ser demitido por justa causa se...*" rather than "*demissão justa causa CLT*"). Each query is paired with 5 PT-BR sources (150–400 words each); sources are synthetic-realistic (generated, not scraped), labeled as such in the dataset's `origem` field — see §6. Queries and sources were generated by a model from a different family than either evaluation engine (Grok), a deliberate separation intended to avoid the party that produces study material sharing a model family with the party that judges it.

Primary engine. `gemini-3.5-flash-lite`, called directly via API (never through a personal-plan chat/CLI interface, which would inject an unobservable system prompt, memory, or personalization) at temperature 0.7 for engine responses and temperature 0 for source transformations, with 3 repetitions per (query, technique) cell. This engine choice is itself the end of a forced substitution chain, recorded here for transparency rather than silently updated: the study's original contract specified `gemini-2.5-flash`, which became unavailable for new API accounts; `gemini-3.6-flash` was tried next but its projected full-run cost ($\approx 5\times$ more expensive) exceeded the study's cost ceiling; `gemini-3.5-flash-lite` was adopted as the final substitution, at the same price point the original contract assumed. We flag this substitution chain explicitly in §6, since a smaller/cheaper "lite" model than originally planned may follow complex transformation instructions less faithfully than a larger model would.

The three evaluation engines. The full benchmark is executed three times, once per engine, on the *same* 525 queries: the primary engine above, `claude-haiku-4-5`, and `gpt-5.6-luna`. The three are drawn from three different model families and are matched by tier — each is its provider's small/economy model of the current generation — rather than by price or parameter count, neither of which is disclosed consistently across providers.

Engine identifiers in this paper are not names we chose: each is the `model_version` string the provider's API returned, recorded in every trace and released with them. Where a provider returns a pinned snapshot, that snapshot is the identifier of record — the Anthropic engine resolves to `claude-haiku-4-5-20251001`, and we write the shorter alias in running text only for readability. A reader checking reproducibility should take the strings in `traces/`, not the ones here.

To isolate the engine variable, **the transformation step is not re-run per engine**: all three engines receive the identical source texts, transformed once by the primary engine. Any difference between the three runs is therefore attributable to the answering/citation step, not to one engine having produced better or worse optimized text than another. The cost of this choice is that the transformations are, in a sense, "native" to one of the three engines; §6 discusses what this does and does not tell us about self-preference.

Two engine-side asymmetries are forced rather than chosen, and are carried into §6 rather than smoothed over: the GPT-5 family does not accept a custom temperature, so `gpt-5.6-luna` runs at its provider default while the other two run at 0.7; and the engines were not collected in a single shared time window (see below).

Execution phases and cost. Pilot and early-phase calls run synchronously against free/low-volume API tiers with backoff; the full run (Phase 3) is required to run in the API's paid batch mode ($\approx$50% cost discount) within a compact collection window (target $\leq$72h) so that no two traces in the final dataset are collected against different model versions of the engine — a risk that a multi-week, free-tier collection window could not rule out. This single-version guarantee holds *per engine* — every trace of a given engine reports one and the same `model_version` — but the three engines were not collected within one shared window, since the study was extended to three engines after the first was already complete;

§6 states what that does and does not threaten. The study's total cost ceiling is US$70 (raised from an initial US$40 when the design was extended from one engine to three); every API call's token usage and dollar cost is logged per-trace in `eval/traces/`, and `src/aggregate.py` sums cost directly from those recorded traces (never recomputed from a price table) to keep the reported cost anchored to what was actually spent.

Parameter	Value
Queries	525 total: 500 evaluation (167/167/166 across health/legal/real-estate) + 25 pitfall
Sources per query	5 (150–400 words, PT-BR)
Techniques	9 (Table 1)
Repetitions per cell	3
Engines (all 525 queries each)	`gemini-3.5-flash-lite`, `claude-haiku-4-5`, `gpt-5.6-luna` (direct API)
Transformation	performed once, reused across all three engines
Engine temperature	0.7 (response) / 0 (transformation); `gpt-5.6-luna` provider default (custom temperature unsu
Cost ceiling	US$70
Collection window	compact and single-`model_version` per engine; not shared across engines

Table 2: Study design parameters, as versioned in `config/study.yaml` (`versao_experimento` = `0.3.0` at the time of writing).

5 Results

All tables in this section are generated by `scripts/csv_to_latex.py` directly from `results/*.csv`, which is itself generated by `src/aggregate.py` from real execution traces in `eval/traces/` — no number in this section is hand-typed. Where a table has not yet been generated (e.g., before Phase 3/4 completion), the corresponding subsection shows an explicit placeholder rather than an omission, so that a reader can distinguish "not run yet" from "ran and found nothing."

5.1 Main results

On the primary engine and the primary query set ($n = 497$ queries with positive baseline visibility), six of the nine techniques move Imp_{pwc} by an amount whose bootstrap CI excludes zero, and the split is lopsided in a direction the original paper does not predict: two are positive and four are negative.

The two positive effects are small. Technical Terms is the largest at a median $+2.6\%$ (CI $[+0.9, +4.3]$) and Cite Sources follows at $+1.8\%$ (CI $[+0.2, +3.9]$). The four negative effects are larger: Keyword Stuffing -9.8% (CI $[-11.7, -5.7]$), Easy-to-Understand -5.4% (CI $[-7.6, -3.6]$), Authoritative -2.4% (CI $[-4.3, -0.3]$) and Unique Words -2.1% (CI $[-4.2, -0.2]$). The remaining three — Quotation Addition ($+0.5\%$), Fluency Optimization ($+0.7\%$) and Statistics Addition (-1.3%) — have CIs that cross zero, and we report them as not distinguishable from no effect rather than as small positive or negative results.

Two observations follow. First, the largest single effect in the whole table is a *penalty*: the worst technique costs about five times more visibility than the best one gains. Second, the ordering does not reproduce the original paper's headline. Of its three strongest techniques — Quotation Addition, Statistics Addition and Cite Sources — only Cite Sources is positive and distinguishable here, and the other two land in the indistinguishable band. §5.5 quantifies that divergence with rank correlation rather than by comparing magnitudes, which the two studies' scales do not permit.

A caveat on reading the sensitivity rows of the table: the `all` rows (`todas` in the released CSV) include queries whose target source was never cited at all, which enter Eq. 3 as an exact 0% (§3.5). On this engine that set is small — 3 of 500 queries — so the two row sets nearly coincide. That is a property of this engine, not of the method, and §5.2 shows it does not hold for the others.

set	technique	metric	n	base.	tech.	median %	CI lo	CI hi	CI excl. 0	mean %	sig. (Holm)
baseline_pos	cite_sources	pwc	497	0.1474	0.1511	1.809	0.2001	3.904	yes (+)	15.35	no
all	cite_sources	pwc	500	0.1465	0.1505	1.740	0.2001	3.796	yes (+)	15.32	sens.
baseline_pos	cite_sources	wc	497	0.1998	0.2056	2.218	0.3898	3.657	yes (+)	10.76	sens.
all	cite_sources	wc	500	0.1986	0.2048	2.175	0.3590	3.657	yes (+)	10.73	sens.
baseline_pos	quotation_addition	pwc	497	0.1474	0.1481	0.5429	-0.6744	2.595	no	11.77	no
all	quotation_addition	pwc	500	0.1465	0.1473	0.5363	-0.6521	2.132	no	11.75	sens.
baseline_pos	quotation_addition	wc	497	0.1998	0.2021	0.3026	-0.0803	1.558	no	7.763	sens.
all	quotation_addition	wc	500	0.1986	0.2011	0.2370	-0.0635	1.558	no	7.747	sens.
baseline_pos	statistics_addition	pwc	497	0.1474	0.1471	-1.347	-3.286	0.7224	no	11.02	no
all	statistics_addition	pwc	500	0.1465	0.1464	-1.324	-3.114	0.5866	no	10.98	sens.
baseline_pos	statistics_addition	wc	497	0.1998	0.2017	0	-2.153	1.436	no	8.769	sens.
all	statistics_addition	wc	500	0.1986	0.2007	0	-2.104	1.275	no	8.734	sens.
baseline_pos	fluency_optimization	pwc	497	0.1474	0.1501	0.6723	-0.3636	2.483	no	10.75	no
all	fluency_optimization	pwc	500	0.1465	0.1493	0.6499	-0.3601	2.313	no	10.71	sens.
baseline_pos	fluency_optimization	wc	497	0.1998	0.2055	0.7321	0	2.292	no	9.719	sens.
all	fluency_optimization	wc	500	0.1986	0.2044	0.5522	0	2.201	no	9.680	sens.
baseline_pos	easy_to_understand	pwc	497	0.1474	0.1379	-5.365	-7.573	-3.571	yes (−)	5.697	yes (−)
all	easy_to_understand	pwc	500	0.1465	0.1376	-5.255	-7.573	-3.459	yes (−)	5.686	sens.
baseline_pos	easy_to_understand	wc	497	0.1998	0.1892	-3.070	-4.589	-1.497	yes (−)	1.506	sens.
all	easy_to_understand	wc	500	0.1986	0.1887	-3.068	-4.553	-1.383	yes (−)	1.503	sens.
baseline_pos	authoritative	pwc	497	0.1474	0.1434	-2.358	-4.307	-0.2585	yes (−)	8.970	no
all	authoritative	pwc	500	0.1465	0.1427	-2.343	-4.193	0	yes (−)	8.934	sens.
baseline_pos	authoritative	wc	497	0.1998	0.1945	-2.175	-3.733	0	yes (−)	4.661	sens.
all	authoritative	wc	500	0.1986	0.1936	-2.120	-3.733	0	no	4.642	sens.
baseline_pos	unique_words	pwc	497	0.1474	0.1424	-2.139	-4.166	-0.2178	yes (−)	1.817	no
all	unique_words	pwc	500	0.1465	0.1416	-2.062	-4.230	-0.2078	yes (−)	1.810	sens.
baseline_pos	unique_words	wc	497	0.1998	0.1927	-1.691	-3.264	-0.5802	yes (−)	0.3775	sens.
all	unique_words	wc	500	0.1986	0.1917	-1.665	-3.249	-0.5139	yes (−)	0.3760	sens.
baseline_pos	technical_terms	pwc	497	0.1474	0.1522	2.553	0.9487	4.282	yes (+)	11.73	yes (+)
all	technical_terms	pwc	500	0.1465	0.1514	2.526	0.9487	4.084	yes (+)	11.70	sens.
baseline_pos	technical_terms	wc	497	0.1998	0.2073	2.546	1.152	4.109	yes (+)	9.871	sens.
all	technical_terms	wc	500	0.1986	0.2062	2.534	1.072	3.969	yes (+)	9.851	sens.
baseline_pos	keyword_stuffing	pwc	497	0.1474	0.1314	-9.843	-11.70	-5.697	yes (−)	-4.019	yes (−)
all	keyword_stuffing	pwc	500	0.1465	0.1307	-9.751	-11.66	-5.590	yes (−)	-4.003	sens.
baseline_pos	keyword_stuffing	wc	497	0.1998	0.1751	-8.569	-11.30	-6.442	yes (−)	-7.724	sens.
all	keyword_stuffing	wc	500	0.1986	0.1741	-8.423	-10.93	-6.409	yes (−)	-7.693	sens.

Table 3: Main results: position-adjusted word count (Imp_{pwc}) and word count (Imp_{wc}) of the target source, baseline vs. technique, with bootstrap 95% CI and relative improvement (mean and median, Eq. 4). Each technique appears twice, once per query set: `baseline_pos` (**primary**) restricts to queries whose target source has positive baseline visibility, and `all` (sensitivity) adds the queries where the source was never cited, which enter Eq. 4 as an exact 0% and pull the median toward zero. Read the primary row as the result; see §3.5. **Paper view:** 12 of 25 columns shown. The complete table is released in `results/tabela_principal.csv`.

5.2 Cross-engine comparison: does the effect depend on who answers?

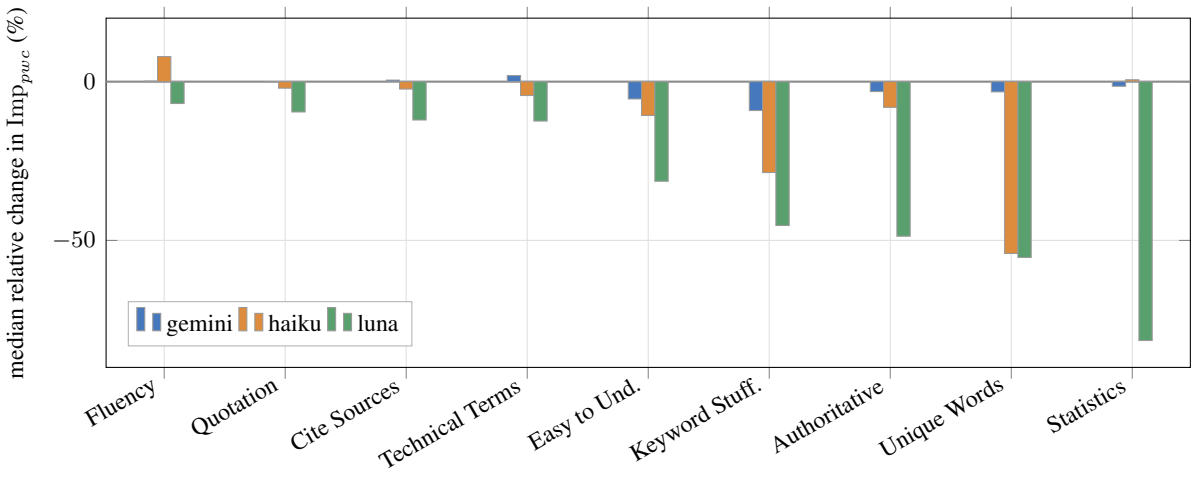

Figure 1: Median relative change in target-source visibility per technique, on each engine, over the primary query set (`baseline_pos`, Imp$_{pwc}$). Techniques are ordered by their worst effect on any engine. Two things are visible at a glance and are the paper's argument: the bars are overwhelmingly below zero, and their heights do not agree across engines — Statistics Addition is negligible on two engines and -81.5% on the third. Exact values, intervals and corrected significance are in Table 4.

This is the central comparison of the study. All three engines answered the same queries with the same already-transformed sources (§4), so a difference between them is a difference in citation behavior, not in the quality of the optimized text. We compare them on a *strictly paired* query set: a query enters only if it is complete on *every* engine (baseline plus all nine techniques, all repetitions present) and the target source occupies the same position on all of them. Completeness is verified per cell rather than per output file, for reasons §6 describes.

Table 4: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Paper view:** rows restricted to `metrica=pwc`; 9 of 28 columns shown. The complete table is released in `results/comparacao_3_engines.csv`.

set	technique	engine	median %	CI lo	CI hi	sig. (Holm)	sig. (BH)	% zeroed
paired	cite_sources	gemini-3.5-flash-lite	1.740	0.2001	3.796	sens.	sens.	0.2012
paired	cite_sources	claude-haiku-4-5	0	-2.604	0	sens.	sens.	6.697
paired	cite_sources	gpt-5.6-luna	-7.259	-12.92	-3.509	sens.	sens.	6.849
paired	quotation_addition	gemini-3.5-flash-lite	0.5363	-0.6521	2.132	sens.	sens.	0.2012
paired	quotation_addition	claude-haiku-4-5	0	-2.361	0	sens.	sens.	5.774
paired	quotation_addition	gpt-5.6-luna	-3.974	-8.520	-0.5781	sens.	sens.	5.936
paired	statistics_addition	gemini-3.5-flash-lite	-1.324	-3.114	0.5866	sens.	sens.	0
paired	statistics_addition	claude-haiku-4-5	0	-3.469	0.0803	sens.	sens.	8.545
paired	statistics_addition	gpt-5.6-luna	-74.56	-82.39	-65.13	sens.	sens.	44.52
paired	fluency_optimization	gemini-3.5-flash-lite	0.6499	-0.3601	2.313	sens.	sens.	0.2012
paired	fluency_optimization	claude-haiku-4-5	5.041	0.8291	11.35	sens.	sens.	4.388
paired	fluency_optimization	gpt-5.6-luna	-3.927	-6.257	0	sens.	sens.	2.283
paired	easy_to_understand	gemini-3.5-flash-lite	-5.255	-7.573	-3.459	sens.	sens.	0.2012
paired	easy_to_understand	claude-haiku-4-5	-7.220	-11.30	-1.385	sens.	sens.	6.928
paired	easy_to_understand	gpt-5.6-luna	-27.35	-31.36	-20.95	sens.	sens.	8.447
paired	authoritative	gemini-3.5-flash-lite	-2.343	-4.193	0	sens.	sens.	0.4024

continued on next page

Table 4 – continued from previous page

set	technique	engine	median %	CI lo	CI hi	sig. (Holm)	sig. (BH)	% zeroed
paired	authoritative	claude-haiku-4-5	-3.846	-8.229	0	sens.	sens.	7.621
paired	authoritative	gpt-5.6-luna	-43.29	-49.07	-36.16	sens.	sens.	21.69
paired	unique_words	gemini-3.5-flash-lite	-2.062	-4.230	-0.2078	sens.	sens.	0.6036
paired	unique_words	claude-haiku-4-5	-46.46	-54.38	-41.30	sens.	sens.	28.87
paired	unique_words	gpt-5.6-luna	-50.02	-54.82	-44.07	sens.	sens.	23.06
paired	technical_terms	gemini-3.5-flash-lite	2.526	0.9487	4.084	sens.	sens.	0.2012
paired	technical_terms	claude-haiku-4-5	-0.7657	-5.769	0	sens.	sens.	7.621
paired	technical_terms	gpt-5.6-luna	-8.906	-11.79	-4.669	sens.	sens.	7.078
paired	keyword_stuffing	gemini-3.5-flash-lite	-9.751	-11.66	-5.590	sens.	sens.	0.2012
paired	keyword_stuffing	claude-haiku-4-5	-25.24	-30.06	-19.62	sens.	sens.	18.01
paired	keyword_stuffing	gpt-5.6-luna	-40.71	-44.44	-33.19	sens.	sens.	17.12
baseline_pos	cite_sources	gemini-3.5-flash-lite	0.4816	-1.216	2.409	no	no	0
baseline_pos	cite_sources	claude-haiku-4-5	-2.329	-5.646	2.169	no	no	5.721
baseline_pos	cite_sources	gpt-5.6-luna	-12.08	-17.13	-7.526	yes (−)	yes (−)	5.721
baseline_pos	quotation_addition	gemini-3.5-flash-lite	0.0055	-1.485	1.591	no	no	0
baseline_pos	quotation_addition	claude-haiku-4-5	-2.051	-5.710	1.594	no	no	4.229
baseline_pos	quotation_addition	gpt-5.6-luna	-9.551	-12.81	-5.194	yes (−)	yes (−)	5.473
baseline_pos	statistics_addition	gemini-3.5-flash-lite	-1.449	-3.560	0.5726	no	no	0
baseline_pos	statistics_addition	claude-haiku-4-5	0.5858	-8.572	4.280	no	no	5.721
baseline_pos	statistics_addition	gpt-5.6-luna	-81.53	-91.78	-73.88	yes (−)	yes (−)	42.29
baseline_pos	fluency_optimization	gemini-3.5-flash-lite	0.1682	-1.377	1.632	no	no	0
baseline_pos	fluency_optimization	claude-haiku-4-5	7.886	2.014	13.18	yes (+)	yes (+)	3.731
baseline_pos	fluency_optimization	gpt-5.6-luna	-6.891	-9.592	-4.514	yes (−)	yes (−)	1.493
baseline_pos	easy_to_understand	gemini-3.5-flash-lite	-5.447	-7.600	-3.683	yes (−)	yes (−)	0
baseline_pos	easy_to_understand	claude-haiku-4-5	-10.66	-15.49	-7.171	yes (−)	yes (−)	5.970
baseline_pos	easy_to_understand	gpt-5.6-luna	-31.37	-34.72	-27.50	yes (−)	yes (−)	6.965
baseline_pos	authoritative	gemini-3.5-flash-lite	-3.132	-4.882	-0.9540	yes (−)	yes (−)	0.2488
baseline_pos	authoritative	claude-haiku-4-5	-8.092	-15.98	-3.878	yes (−)	yes (−)	6.716
baseline_pos	authoritative	gpt-5.6-luna	-48.68	-55.75	-43.39	yes (−)	yes (−)	19.65
baseline_pos	unique_words	gemini-3.5-flash-lite	-3.218	-5.183	-1.001	yes (−)	yes (−)	0
baseline_pos	unique_words	claude-haiku-4-5	-54.12	-65.80	-46.73	yes (−)	yes (−)	26.12
baseline_pos	unique_words	gpt-5.6-luna	-55.37	-60.05	-51.26	yes (−)	yes (−)	20.65
baseline_pos	technical_terms	gemini-3.5-flash-lite	1.920	0.0067	3.814	no	no	0
baseline_pos	technical_terms	claude-haiku-4-5	-4.371	-8.163	-0.0948	no	no	5.473
baseline_pos	technical_terms	gpt-5.6-luna	-12.41	-16.98	-9.483	yes (−)	yes (−)	6.219
baseline_pos	keyword_stuffing	gemini-3.5-flash-lite	-9.103	-11.53	-5.459	yes (−)	yes (−)	0.2488
baseline_pos	keyword_stuffing	claude-haiku-4-5	-28.62	-34.36	-23.13	yes (−)	yes (−)	15.17
baseline_pos	keyword_stuffing	gpt-5.6-luna	-45.31	-50.11	-41.07	yes (−)	yes (−)	15.67

Direction agreement. We report, per technique, the sign of the median relative improvement on each engine and whether all engines agree. Two conventions are fixed in advance and never relaxed: the denominator is *all nine* techniques of the study, and a technique whose effect is indeterminate on any engine counts as non-agreement — the denominator is never reduced to the subset that happened to be measurable. We further separate the two ways engines can disagree, since they are different claims in prose: *inversion* (the effect is positive on one engine and negative on another) and *attenuation* (the effect is present on one and null on another).

On the primary query set the effective n is 402 and is identical on all three engines by construction (§3.5), so every number below is computed over exactly the same queries.

The effect of a technique depends on which engine answers. The three engines agree on the direction of the effect for **four of the nine techniques** (44.4%). The four are Easy-to-Understand (-5.4% / -10.7% / -31.4% for gemini / haiku / luna), Authoritative (-3.1% / -8.1% / -48.7%), Unique Words (-3.2% / -54.1% / -55.4%) and Keyword Stuffing (-9.1% / -28.6% / -45.3%). All four are *negative*, and all four have CIs excluding zero on all three engines: what transfers reliably across engines in this benchmark is which techniques *hurt*, not which help.

set	metric	technique	gemini med. %	haiku med. %	luna med. %	gemini CI	haiku CI	luna CI	agree	divergence	sig. all	n eff.
paired	pwc	cite_sources	1.740	0	-7.259	yes (+)	no	yes (−)	no	inversion	no	468
paired	pwc	quotation_addition	0.5363	0	-3.974	no	no	yes (−)	no	inversion	no	469
paired	pwc	statistics_addition	-1.324	0	-74.56	no	no	yes (−)	no	attenuation	no	467
paired	pwc	fluency_optimization	0.6499	5.041	-3.927	no	yes (+)	no	no	inversion	no	454
paired	pwc	easy_to_understand	-5.255	-7.220	-27.35	yes (−)	yes (−)	yes (−)	yes	—	yes	466
paired	pwc	authoritative	-2.343	-3.846	-43.29	yes (−)	no	yes (−)	yes	—	no	475
paired	pwc	unique_words	-2.062	-46.46	-50.02	yes (−)	yes (−)	yes (−)	yes	—	yes	491
paired	pwc	technical_terms	2.526	-0.7657	-8.906	yes (+)	no	yes (−)	no	inversion	no	471
paired	pwc	keyword_stuffing	-9.751	-25.24	-40.71	yes (−)	yes (−)	yes (−)	yes	—	yes	479
paired	wc	cite_sources	2.175	0	-7.461	yes (+)	no	yes (−)	no	inversion	no	468
paired	wc	quotation_addition	0.2370	0	-3.520	no	no	yes (−)	no	inversion	no	469
paired	wc	statistics_addition	0	0	-72.28	no	no	yes (−)	no	attenuation	no	467
paired	wc	fluency_optimization	0.5522	6.660	-2.846	no	yes (+)	no	no	inversion	no	454
paired	wc	easy_to_understand	-3.068	-6.654	-26.30	yes (−)	yes (−)	yes (−)	yes	—	yes	466
paired	wc	authoritative	-2.120	-5.682	-41.28	no	yes (−)	yes (−)	yes	—	no	475
paired	wc	unique_words	-1.665	-45.94	-48.41	yes (−)	yes (−)	yes (−)	yes	—	yes	491
paired	wc	technical_terms	2.534	-2.291	-8.004	yes (+)	no	yes (−)	no	inversion	no	471
paired	wc	keyword_stuffing	-8.423	-26.81	-39.22	yes (−)	yes (−)	yes (−)	yes	—	yes	479
baseline_pos	pwc	cite_sources	0.4816	-2.329	-12.08	no	no	yes (−)	no	inversion	no	402
baseline_pos	pwc	quotation_addition	0.0055	-2.051	-9.551	no	no	yes (−)	no	inversion	no	402
baseline_pos	pwc	statistics_addition	-1.449	0.5858	-81.53	no	no	yes (−)	no	inversion	no	402
baseline_pos	pwc	fluency_optimization	0.1682	7.886	-6.891	no	yes (+)	yes (−)	no	inversion	no	402
baseline_pos	pwc	easy_to_understand	-5.447	-10.66	-31.37	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	pwc	authoritative	-3.132	-8.092	-48.68	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	pwc	unique_words	-3.218	-54.12	-55.37	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	pwc	technical_terms	1.920	-4.371	-12.41	yes (+)	yes (−)	yes (−)	no	inversion	yes	402
baseline_pos	pwc	keyword_stuffing	-9.103	-28.62	-45.31	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	wc	cite_sources	0.6922	-0.8432	-11.33	no	no	yes (−)	no	inversion	no	402
baseline_pos	wc	quotation_addition	0	-0.5784	-8.547	no	no	yes (−)	no	attenuation	no	402
baseline_pos	wc	statistics_addition	-0.1738	0.7699	-80.09	no	no	yes (−)	no	inversion	no	402
baseline_pos	wc	fluency_optimization	0.3651	9.161	-6.529	no	yes (+)	yes (−)	no	inversion	no	402
baseline_pos	wc	easy_to_understand	-3.247	-10.97	-30.66	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	wc	authoritative	-2.805	-9.352	-47.23	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	wc	unique_words	-2.051	-52.79	-53.98	yes (−)	yes (−)	yes (−)	yes	—	yes	402
baseline_pos	wc	technical_terms	1.814	-5.965	-12.28	yes (+)	yes (−)	yes (−)	no	inversion	yes	402
baseline_pos	wc	keyword_stuffing	-8.575	-30.58	-43.84	yes (−)	yes (−)	yes (−)	yes	—	yes	402

Table 5: Direction-of-effect agreement across engines, per technique: the sign of the median relative improvement on each engine and whether all engines agree. The denominator is all nine techniques of the study and an indeterminate result counts as non-agreement — the denominator is never reduced to the measurable subset. Disagreements are further split into *inversion* (opposite signs) and *attenuation* (effect present on one engine, null on another), since the two support different claims. The per-engine significance columns are uncorrected bootstrap intervals; under Holm correction the only inversion with both arms surviving is Fluency Optimization (§5.2). **Paper view:** 13 of 18 columns shown. The complete table is released in `results/concordancia_engines.csv`.

set	metric	engine A	engine B	ρ	exact p	crit. $\lvert\rho\rvert$	$\rho \neq 0$
paired	pwc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6781	0.0522	0.6950	no
paired	pwc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no
paired	pwc	claude-haiku-4-5	gpt-5.6-luna	0.5933	0.1010	0.6950	no
paired	wc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6781	0.0522	0.6950	no
paired	wc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no
paired	wc	claude-haiku-4-5	gpt-5.6-luna	0.5933	0.1010	0.6950	no
baseline_pos	pwc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6667	0.0589	0.7000	no
baseline_pos	pwc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5500	0.1328	0.7000	no
baseline_pos	pwc	claude-haiku-4-5	gpt-5.6-luna	0.4833	0.1938	0.7000	no
baseline_pos	wc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6167	0.0857	0.7000	no
baseline_pos	wc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no
baseline_pos	wc	claude-haiku-4-5	gpt-5.6-luna	0.4833	0.1938	0.7000	no

Table 6: Pairwise Spearman rank correlation between engines over the nine-technique vector of median relative improvements. The p is exact — the permutation null over the 9! orderings is enumerated in full, since the asymptotic approximation is not valid at $n = 9$. **No pair reaches significance**: the critical $\lvert\rho\rvert$ is 0.700 (lower where ties in the ranks shrink the null), above every ρ observed here. These correlations are reported for completeness and support no claim about ordering; the per-technique effects and their intervals (Table 5) are what this study leans on. **Paper view:** 8 of 10 columns shown. The complete table is released in `results/spearman_engines.csv`.

The agreement rate is itself sensitive to how many engines are asked. The same measurement over only the first two engines gives 66.7% (6 of 9); adding a third engine reduces it to 44.4%. A two-engine cross-validation of the kind originally planned for this study would therefore have reported substantially more agreement than three engines support — which is an argument about the design of GEO evaluations, not only about our numbers.

Multiplicity, and what survives it. Before naming any single technique as the study's inversion, the family of tests has to be accounted for. We pre-specify the primary family as the nine techniques on every engine, in the primary query set and primary metric — $9 \times 3 = 27$ tests — and apply Holm–Bonferroni to it. Holm rather than Benjamini–Hochberg because a single false positive in the main table is what would invert the paper's reading, so the quantity to control is the probability of *any* false positive, not their proportion. The p values come from the same bootstrap distribution that produced the intervals, so the corrected test and the interval are two views of one resampling, not two different tests. Breakdowns by sector and by position are declared exploratory and are *not* corrected — not because the family would grow, but because they were not pre-specified as confirmatory; every row of those tables carries that label in the released CSV.

Eighteen of the 27 primary tests survive correction, against 20 before it. The two that fall are consequential, and we report the fall rather than the uncorrected reading.

The choice of correction does not drive this. Benjamini–Hochberg, the natural alternative if one reads the table as a list of candidates to investigate rather than as confirmatory tests, is reported in the same table and selects *exactly the same 18* — no test changes status between the two, including the two that fall. We report both so that the choice of correction cannot be read as a choice of result, which is the same reason §5.5 refuses to switch estimators.

The inversion that survives is Fluency Optimization, not Technical Terms. At the uncorrected level, Technical Terms looks like this study's cleanest result: positive on gemini ($+1.9\%$), negative on haiku (-4.4%) and luna (-12.4%), every interval excluding zero. It does not survive. Its raw p on gemini and haiku is 0.039 and 0.043 — just inside the threshold — and Holm carries both to 0.35. Only the luna arm ($p_{\text{Holm}} = 0.003$) remains. This is robust to how the family is drawn: correcting within each engine separately ($m = 9$), the most lenient defensible choice, still leaves gemini at 0.19 and haiku at 0.17. After correction, Technical Terms is a technique with one significant negative effect and two indeterminate ones — an attenuation, not an inversion.

What does survive is Fluency Optimization, and it is a better example than the one it replaces. It is significantly *positive* on haiku ($+7.9\%$, $p_{\text{Holm}} = 0.031$) and significantly *negative* on luna (-6.9%, $p_{\text{Holm}} = 0.003$), with gemini indeterminate ($+0.2\%$). Two engines, opposite signs, both arms surviving a correction applied across the whole primary grid. And the technique is the least exotic of the nine: it does not add statistics, fabricate citations or stuff keywords — it rewrites the source to read better. That the mildest transformation in the catalogue helps one engine and hurts another, both distinguishably, is a sharper statement of engine-dependence than the inversion we would have reported without correcting.

We flag one asymmetry that applies to whichever inversion is named: on gemini — the engine whose family produced the transformed text — self-preference is a live alternative explanation for any positive arm, and this design cannot rule it out. It does not touch the negative arms, since haiku and luna had no hand in producing the text they penalize. In the surviving inversion the positive arm is on haiku, which did not transform the text, so self-preference does not explain it.

The shape of the effect, not only its sign. The negative arms are partly threshold effects rather than uniform shrinkage, and this is measured. Technical Terms removes the target source from the answer entirely in 0.0% of queries on gemini, 5.5% on haiku and 6.2% on luna; for Unique Words the same figure reaches 26.1% on haiku and 20.6% on luna against 0.0% on gemini, and for Statistics Addition 42.3% on luna. On a substantial minority of queries the optimized source is not cited less — it stops being cited. This suggests the effect decomposes into two stages, the probability that a source is cited

at all and its share of the answer given that it is cited, which a percentage change in share conflates. We report the zeroing rate alongside the median so the two are visible separately, but we do not fit the two-stage model here; §6 records it as the analysis this dataset most obviously supports and this paper does not perform.

Why fluent prose should raise one engine's use of a source and lower another's is a question about the engines' internal preferences, which this design observes from the outside and cannot answer.

We are deliberately strict throughout. Five techniques show a sign disagreement in the medians; most do not survive as inversions once significance — corrected significance — is demanded on both arms. Statistics Addition illustrates the distinction: its median is -81.5% on luna, the largest single effect anywhere in this study and comfortably surviving correction, while gemini (-1.4%) and haiku ($+0.6\%$) are indeterminate. The honest description is that one engine reacts catastrophically where the others show nothing, not that the engines disagree in direction.

Magnitude, and the mechanism behind it. Effect sizes are not comparable in scale across engines: on the techniques where all three agree, luna's median response is roughly 5–17$\times$ gemini's, with haiku in between. The citation-zeroed column identifies the mechanism, and it is not that the optimized source is cited a little less. It is that the source *stops being cited at all*. Among queries whose target source was cited in the baseline, Unique Words removes it entirely from 26.1% of answers on haiku and 20.6% on luna, and Statistics Addition removes it from 42.3% of answers on luna. On gemini the same figure never reaches 0.3% for any technique: that engine essentially never drops a source it had already cited, whatever is done to the text.

This also explains why the two query sets in the table diverge on two engines and not on the third (§3.5): a technique that pushes sources out of the answer generates exactly the degenerate zero-baseline queries that pull a pooled median toward zero.

Rank agreement. Spearman correlation over the nine-technique vectors is $\rho = 0.67$ (gemini–haiku), 0.55 (gemini–luna) and 0.48 (haiku–luna). **None of the three reaches significance**: the exact two-sided p values, from exhaustive permutation of the $9!$ orderings, are 0.059, 0.133 and 0.194 against a critical $|\rho|$ of 0.700.

We stress what this does and does not establish. It is not evidence that the engines rank the techniques alike — no pair does so distinguishably from chance. Neither is it evidence that they rank them differently: failing to reject the null is not accepting it, and nine points cannot settle the question in either direction. The claim of this section rests on the per-technique effects under correction (Table 5): direction agreement on only four of nine, and the Fluency Optimization inversion with both arms surviving Holm — not on the rank correlations, which we report for completeness and lean on for nothing.

5.3 Breakdown by sector

Query volume per sector is comparable and adequate ($n = 166$ health, 167 legal, 164 real estate on the primary set), so sector differences here are not an artifact of one sector being thinly sampled.

Health, the YMYL sector, is the *least* responsive of the three. Only two of nine techniques move visibility distinguishably there — Technical Terms $+3.2\%$ (CI $[+0.4, +6.8]$) and Keyword Stuffing -5.9% (CI $[-10.4, -3.4]$) — against three in legal and four in real estate. This runs against the intuition that a YMYL topic would make an engine *more* reactive to signals of authority: Authoritative is not distinguishable from zero in any of the three sectors, health included.

Only one effect is stable across all three sectors, and it is a penalty: Keyword Stuffing is negative and distinguishable everywhere, and it is weakest precisely in health (-5.9%, versus -10.9% legal and -11.1% real estate). Technical Terms, the technique with the largest median gain on this engine, is positive and distinguishable in health and real estate but not in legal ($+1.1\%$, CI crossing zero) — a reminder that a single headline number per technique conceals sector-level variation of the same order as the effect itself.

sector	technique	n	base.	tech.	median %	CI lo	CI hi	CI excl. 0	mean %
health	cite_sources	166	0.1477	0.1491	0.8963	-1.619	5.357	no	26.27
health	quotation_addition	166	0.1477	0.1479	0.0854	-2.718	3.761	no	23.20
health	statistics_addition	166	0.1477	0.1529	-0.1875	-3.804	3.122	no	26.01
health	fluency_optimization	166	0.1477	0.1475	0	-2.872	2.291	no	15.13
health	easy_to_understand	166	0.1477	0.1418	-3.735	-7.561	0	no	23.66
health	authoritative	166	0.1477	0.1465	-2.053	-4.866	2.796	no	23.14
health	unique_words	166	0.1477	0.1438	-0.0781	-3.415	2.033	no	3.009
health	technical_terms	166	0.1477	0.1526	3.205	0.4254	6.787	yes (+)	15.63
health	keyword_stuffing	166	0.1477	0.1355	-5.878	-10.35	-3.401	yes (−)	5.181
legal	cite_sources	167	0.1429	0.1471	1.853	-0.3376	4.950	no	9.170
legal	quotation_addition	167	0.1429	0.1456	0.6809	-1.694	2.826	no	7.538
legal	statistics_addition	167	0.1429	0.1384	-3.484	-6.718	0	yes (−)	0.8321
legal	fluency_optimization	167	0.1429	0.1471	1.671	-0.4702	5.156	no	9.422
legal	easy_to_understand	167	0.1429	0.1297	-6.971	-13.91	-3.634	yes (−)	-4.699
legal	authoritative	167	0.1429	0.1375	-2.947	-6.769	0.3179	no	3.153
legal	unique_words	167	0.1429	0.1382	-2.790	-6.483	0	no	2.542
legal	technical_terms	167	0.1429	0.1456	1.141	-1.608	5.144	no	8.515
legal	keyword_stuffing	167	0.1429	0.1249	-10.92	-16.91	-4.786	yes (−)	-9.065
real estate	cite_sources	164	0.1516	0.1572	2.097	-0.4852	6.825	no	10.59
real estate	quotation_addition	164	0.1516	0.1509	0.6961	-1.132	4.452	no	4.511
real estate	statistics_addition	164	0.1516	0.1501	-0.1359	-3.560	2.498	no	6.220
real estate	fluency_optimization	164	0.1516	0.1557	0.7109	-1.843	3.541	no	7.667
real estate	easy_to_understand	164	0.1516	0.1424	-4.957	-8.663	-1.953	yes (−)	-1.896
real estate	authoritative	164	0.1516	0.1463	-2.331	-5.325	0.2637	no	0.5486
real estate	unique_words	164	0.1516	0.1451	-4.072	-6.770	-0.3960	yes (−)	-0.1277
real estate	technical_terms	164	0.1516	0.1586	2.874	0.8134	5.847	yes (+)	11.04
real estate	keyword_stuffing	164	0.1516	0.1339	-11.07	-13.82	-5.475	yes (−)	-8.192

Table 7: Breakdown by sector (health / legal / real estate). Health is a YMYL (Your-Money-Your-Life) sector; sector-level behavior differences are a target finding of this study. Query sets as in Table 3: `baseline_pos` is primary. **Paper view:** rows restricted to `conjunto=baseline_pos`, `metrica=pwc`; 10 of 26 columns shown. The complete table is released in `results/quebra_por_setor.csv`.

5.4 Breakdown by target-source position

Table 8: Breakdown by target-source position (rank 1–5), alongside the corresponding values from Table 2 of the original GEO paper [Aggarwal et al., 2024]. Query sets as in Table 3: `baseline_pos` is primary. **Paper view:** rows restricted to `conjunto=baseline_pos`, `metrica=pwc`; 9 of 28 columns shown. The complete table is released in `results/quebra_por_posicao.csv`.

rank	technique	n	median %	CI lo	CI hi	CI excl. 0	paper %	replicates
1	cite_sources	110	0.0310	-2.350	2.987	no	-30.30	no
1	quotation_addition	110	-2.084	-4.018	0.6440	no	-22.90	yes
1	statistics_addition	110	-0.0043	-2.991	2.381	no	-20.60	yes
1	fluency_optimization	110	0.3162	-1.789	3.070	no	-2.000	no
1	easy_to_understand	110	-4.681	-6.817	-0.5947	yes (−)	n/d	n/a
1	authoritative	110	-5.013	-6.769	-2.789	yes (−)	-6.000	yes
1	unique_words	110	-4.862	-7.574	0.1566	no	n/d	n/a
1	technical_terms	110	1.641	-1.758	5.322	no	n/d	n/a
1	keyword_stuffing	110	-5.010	-9.996	-1.459	yes (−)	n/d	n/a
2	cite_sources	80	0.6289	-5.939	7.741	no	2.500	yes
2	quotation_addition	80	3.323	-1.666	6.135	no	-7.000	no
2	statistics_addition	80	-1.106	-5.899	2.726	no	-3.900	yes
2	fluency_optimization	80	1.011	-2.844	4.665	no	5.200	yes
2	easy_to_understand	80	-3.061	-9.309	2.190	no	n/d	n/a
2	authoritative	80	-3.483	-8.698	2.188	no	4.100	no
2	unique_words	80	0.5962	-5.939	3.946	no	n/d	n/a
2	technical_terms	80	1.904	-2.826	5.754	no	n/d	n/a
2	keyword_stuffing	80	-11.03	-14.83	-3.798	yes (−)	n/d	n/a
3	cite_sources	89	2.837	1.264	9.143	yes (+)	20.40	yes
3	quotation_addition	89	4.511	0	10.53	no	3.500	yes
3	statistics_addition	89	-1.312	-3.782	4.316	no	8.100	no
3	fluency_optimization	89	1.427	-0.8984	5.964	no	3.600	yes
3	easy_to_understand	89	-8.953	-14.71	-2.518	yes (−)	n/d	n/a
3	authoritative	89	-0.2948	-4.620	4.574	no	-0.6000	yes
3	unique_words	89	-1.372	-5.415	1.230	no	n/d	n/a
3	technical_terms	89	5.043	1.682	9.198	yes (+)	n/d	n/a
3	keyword_stuffing	89	-17.18	-23.95	-11.82	yes (−)	n/d	n/a
4	cite_sources	113	1.228	-1.625	6.825	no	15.50	yes
4	quotation_addition	113	1.776	-2.884	6.880	no	25.10	yes
4	statistics_addition	113	-3.021	-6.856	3.061	no	10.00	no
4	fluency_optimization	113	0	-4.350	4.194	no	-4.400	no
4	easy_to_understand	113	-10.47	-15.39	-5.611	yes (−)	n/d	n/a
4	authoritative	113	-2.343	-10.04	3.017	no	12.60	no
4	unique_words	113	-3.713	-8.147	0	no	n/d	n/a
4	technical_terms	113	2.240	-2.751	5.672	no	n/d	n/a
4	keyword_stuffing	113	-6.762	-17.13	-0.1167	yes (−)	n/d	n/a
5	cite_sources	105	3.454	-1.619	7.646	no	115.1	yes
5	quotation_addition	105	-0.6744	-4.920	2.595	no	99.70	no
5	statistics_addition	105	-4.586	-9.486	1.814	no	97.90	no
5	fluency_optimization	105	0.6723	-2.425	7.769	no	2.200	yes
5	easy_to_understand	105	-2.786	-6.477	3.027	no	n/d	n/a
5	authoritative	105	0.4740	-3.096	4.244	no	6.100	yes
5	unique_words	105	-1.378	-6.269	2.247	no	n/d	n/a
5	technical_terms	105	1.529	-0.0871	6.928	no	n/d	n/a
5	keyword_stuffing	105	-7.156	-12.22	-2.589	yes (−)	n/d	n/a

The original paper's headline position-level finding (Aggarwal et al. 2024, their Table 2) is a steep gradient: Cite Sources raised the visibility of rank-5 (worst-placed) sources by 115.1% while *lowering* rank-1 visibility by 30.3%. **We do not replicate that gradient.** For Cite Sources our medians are $+0.0\%$, $+0.6\%$, $+2.8\%$, $+1.2\%$ and $+3.5\%$ across ranks 1–5, and only the rank-3 value has a CI excluding zero (CI $[+1.3, +9.1]$). The rank-5 point estimate is $+3.5\%$ against the original's $+115.1\%$,

and its CI ($[-1.6, +7.6]$) includes zero.

Sign agreement per position is therefore a misleading summary on its own: Cite Sources matches the original's direction at four of five ranks, but at magnitudes roughly one to two orders smaller and, at four of those five ranks, indistinguishable from no effect at all. We report the sign comparison in the table for completeness and rest no claim on it.

What does show a position pattern is the penalty side. Keyword Stuffing is negative and distinguishable at *every* rank, peaking at rank 3 (-17.2%, CI $[-24.0, -11.8]$), and Easy-to-Understand is distinguishably negative at ranks 1, 3 and 4. The strongest effects of position in this benchmark act on techniques the original paper reports as helpful or neutral, not on the one it highlights.

5.5 Comparison with the original paper

set	technique	paper PAWC	paper Δ%	our median %	CI lo	CI hi	n	CI excl. 0	replicates
baseline_pos	cite_sources	24.60	27.46	1.809	0.2001	3.904	497	yes (+)	yes
baseline_pos	quotation_addition	27.20	40.93	0.5429	-0.6744	2.595	497	no	yes
baseline_pos	statistics_addition	25.20	30.57	-1.347	-3.286	0.7224	497	no	no
baseline_pos	fluency_optimization	24.70	27.98	0.6723	-0.3636	2.483	497	no	yes
baseline_pos	easy_to_understand	22.00	13.99	-5.365	-7.573	-3.571	497	yes (−)	no
baseline_pos	authoritative	21.30	10.36	-2.358	-4.307	-0.2585	497	yes (−)	no
baseline_pos	unique_words	20.50	6.218	-2.139	-4.166	-0.2178	497	yes (−)	no
baseline_pos	technical_terms	22.70	17.62	2.553	0.9487	4.282	497	yes (+)	yes
baseline_pos	keyword_stuffing	17.70	-8.290	-9.843	-11.70	-5.697	497	yes (−)	yes
all	cite_sources	24.60	27.46	1.740	0.2001	3.796	500	yes (+)	yes
all	quotation_addition	27.20	40.93	0.5363	-0.6521	2.132	500	no	yes
all	statistics_addition	25.20	30.57	-1.324	-3.114	0.5866	500	no	no
all	fluency_optimization	24.70	27.98	0.6499	-0.3601	2.313	500	no	yes
all	easy_to_understand	22.00	13.99	-5.255	-7.573	-3.459	500	yes (−)	no
all	authoritative	21.30	10.36	-2.343	-4.193	0	500	yes (−)	no
all	unique_words	20.50	6.218	-2.062	-4.230	-0.2078	500	yes (−)	no
all	technical_terms	22.70	17.62	2.526	0.9487	4.084	500	yes (+)	yes
all	keyword_stuffing	17.70	-8.290	-9.751	-11.66	-5.590	500	yes (−)	yes

Table 9: Direction-of-effect comparison against the original paper's Table 1 (PAWC-Overall, Aggarwal et al. 2024) and Spearman rank correlation. Absolute scales are not comparable across studies (different engines/metrics normalization) — only sign and ordering are. Reported for both query sets: the Spearman ρ differs between them because medians pinned at exactly 0% in the full set become ties and receive averaged ranks; `baseline_pos` is primary. **Paper view:** 10 of 11 columns shown. The complete table is released in `results/comparacao_paper.csv`.

GEO-PTBR reproduces the original paper's *direction* of effect for five of the nine techniques (Cite Sources, Quotation Addition, Fluency Optimization, Technical Terms and Keyword Stuffing) and diverges on four (Statistics Addition, Easy-to-Understand, Authoritative and Unique Words). All four divergences share a shape: the original reports them as positive — between $+6.2\%$ and $+30.6\%$ relative to its baseline — and we measure them as negative, three of the four distinguishably so.

Rank correlation with the original's Table 1 ordering is $\rho = +0.62$ by median and $\rho = +0.80$ by mean.

Neither figure licenses a claim that the ordering is preserved, and the median-based one is not statistically distinguishable from chance. With nine techniques the permutation null can be enumerated exhaustively ($9!$ orderings), so we report an exact two-sided p rather than an asymptotic approximation, which is not valid at this n. The critical value is $|\rho| = 0.700$; our median-based $\rho = +0.62$ gives $p = 0.086$, and only the mean-based $\rho = +0.80$ reaches significance ($p = 0.014$).

A note on that threshold, since printed tables for $n = 9$ often list 0.683: we define the critical value as the smallest attainable $|\rho|$ whose exact p does not exceed 0.05, and 0.683 fails that test by a hair — its exact two-sided p is 0.0503. Nothing here turns on the choice; 0.62 falls short of either. We state the convention only so the number can be checked against the enumeration rather than against a table.

This is an uncomfortable pairing and we report it rather than resolve it in our favour. The statistic we

designated as primary in advance, for reasons independent of this comparison (§3.5), does not support an ordering claim; the statistic we designated as unreliable does. Choosing the mean here would be selecting the estimator by its result. We therefore make no claim that the technique ordering transfers from English to PT-BR: at $n = 9$ this study is simply underpowered to establish it either way, and any future replication comparing technique rankings should treat nine points as a descriptive summary, not a test. A further reason for caution is specific to ordering claims: how much a technique changes the source's length rank-predicts its effect (§5.7), so any per-technique ordering here is partly an ordering by a mechanical property of the rewrite.

What the data *does* support is a claim about level rather than order, and it does not depend on the correlation at all. The original reports its top techniques at $+27$ to $+41\%$ over baseline; our largest positive effect on any engine is $+2.6\%$, four of nine techniques are negative with confidence intervals excluding zero, and the largest effect measured anywhere in this study is a penalty of -81.5%. Absolute magnitudes are not comparable across studies by construction (§3.5), so we do not compare $+40\%$ against $+2.6\%$ as quantities. The comparison that survives is qualitative and, for a practitioner, the decisive one: in the original setting these techniques are a set of reliable gains, and in ours they are mostly not gains at all.

5.6 Citation induction on pitfall queries

technique	metric	n_pegadinhas	baseline_total_mean	baseline_ci_lo	baseline_ci_hi	tecnica_total_mean	tecnica_ci_lo	tecnica_ci_hi	delta_mean	delta_ci_lo	delta_ci_hi	inducao
cite_sources	pwc	25	0	0	0	0.0533	0	0.1333	0.0533	0	0.1333	no
cite_sources	wc	25	0	0	0	0.0533	0	0.1333	0.0533	0	0.1333	no
quotation_addition	pwc	25	0	0	0	0.1972	0.0620	0.3544	0.1972	0.0620	0.3544	yes
quotation_addition	wc	25	0	0	0	0.2271	0.0750	0.3956	0.2271	0.0750	0.3956	yes
statistics_addition	pwc	25	0	0	0	0.3150	0.1578	0.4798	0.3150	0.1578	0.4798	yes
statistics_addition	wc	25	0	0	0	0.3724	0.1932	0.5560	0.3724	0.1932	0.5560	yes
fluency_optimization	pwc	25	0	0	0	0.0385	0	0.1037	0.0385	0	0.1037	no
fluency_optimization	wc	25	0	0	0	0.0449	0	0.1165	0.0449	0	0.1165	no
easy_to_understand	pwc	25	0	0	0	0.2079	0.0896	0.3456	0.2079	0.0896	0.3456	yes
easy_to_understand	wc	25	0	0	0	0.2469	0.1067	0.4069	0.2469	0.1067	0.4069	yes
authoritative	pwc	25	0	0	0	0.2140	0.0765	0.3768	0.2140	0.0765	0.3768	yes
authoritative	wc	25	0	0	0	0.2335	0.0800	0.4000	0.2335	0.0800	0.4000	yes
unique_words	pwc	25	0	0	0	0	0	0	0	0	0	no
unique_words	wc	25	0	0	0	0	0	0	0	0	0	no
technical_terms	pwc	25	0	0	0	0.0537	0	0.1366	0.0537	0	0.1366	no
technical_terms	wc	25	0	0	0	0.0667	0	0.1733	0.0667	0	0.1733	no
keyword_stuffing	pwc	25	0	0	0	0.3148	0.1524	0.4851	0.3148	0.1524	0.4851	yes
keyword_stuffing	wc	25	0	0	0	0.3510	0.1771	0.5301	0.3510	0.1771	0.5301	yes

Table 10: Citation induction on pitfall queries (unanswerable-by-design): total $\mathrm{Imp}_{wc}/\mathrm{Imp}_{pwc}$ across all 5 sources, by technique. A value significantly above zero flags a technique that induced the engine to cite a source that does not answer the query.

Pitfall queries are, by design, unanswerable from the five given sources; baseline citation visibility on them should be $\approx$0, and any technique that raises it above zero has induced the engine to cite a source that does not actually answer the question — a form of hallucination *induced* by the optimization technique itself, not merely tolerated by the baseline pipeline.

The control works as designed: across all 25 pitfall queries the baseline total visibility is exactly 0 (CI $[0, 0]$) for every technique. The measurement pipeline does not hallucinate citations on its own.

The techniques do. **Five of the nine induce citation of a source that cannot answer the query**, with bootstrap CIs excluding zero: Statistics Addition ($\Delta = +0.315$, CI $[+0.158, +0.480]$), Keyword Stuffing ($+0.315$, CI $[+0.152, +0.485]$), Authoritative ($+0.214$, CI $[+0.077, +0.377]$), Easy-to-Understand ($+0.208$, CI $[+0.090, +0.346]$) and Quotation Addition ($+0.197$, CI $[+0.062, +0.354]$). The remaining four — Cite Sources, Fluency Optimization, Technical Terms and Unique Words — are not distinguishable from zero, and Unique Words induces nothing at all ($\Delta = 0$ exactly).

This deserves emphasis because it inverts the usual framing of these techniques as neutral formatting choices. Statistics Addition is the clearest case: instructed to add statistics to a source that does not address the question, the transformation manufactures figures about the queried entity, and the engine then cites that source as though it answered. The technique does not make a source more visible for what it says; it makes it say something it did not. On a YMYL query — and a third of this benchmark is health — that is a safety failure, not an optimization result.

Note the disjunction with §5.1: three of these five (Statistics Addition, Authoritative, Easy-to-Understand) are useless or harmful for genuine visibility on answerable queries while being among the most effective at inducing a citation where none is warranted.

5.7 A confound we can measure: transformations change source length

Visibility is a share of the generated answer attributed to a source. A transformation that leaves the source with more — or less — usable material can move that share without the engine having judged the source better or worse. Until this is measured, "Keyword Stuffing hurts" and "Keyword Stuffing shortens the source by 12%" are indistinguishable claims. We measure it, and the news is not comfortable.

technique	delta_comprimento_pct	gemini med. %	haiku med. %	luna med. %
cite_sources	6.725	0.4816	-2.329	-12.08
quotation_addition	3.361	0.0055	-2.051	-9.551
statistics_addition	-6.061	-1.449	0.5858	-81.53
fluency_optimization	7.310	0.1682	7.886	-6.891
easy_to_understand	-14.41	-5.447	-10.66	-31.37
authoritative	-13.51	-3.132	-8.092	-48.68
unique_words	-9.353	-3.218	-54.12	-55.37
technical_terms	-0.7979	1.920	-4.371	-12.41
keyword_stuffing	-12.43	-9.103	-28.62	-45.31

Table 11: Length change of the transformed source per technique (median over the 525 queries, in words, against the original source), alongside the median visibility change each engine measures. The transformation is performed once and reused across engines, so the length column is engine-independent. Across the nine techniques, length change and visibility change are strongly rank-correlated: $\rho = +0.80$ ($p = 0.014$) on gemini, $+0.73$ ($p = 0.031$) on haiku and $+0.67$ ($p = 0.059$) on luna, exact permutation p with critical $|\rho| = 0.700$. See §5.7 for what this does and does not confound.

The nine techniques change source length by between -14.4% (Easy-to-Understand) and $+7.3\%$ (Fluency Optimization). Across those nine points, length change and visibility change are strongly rank-correlated on every engine: $\rho = +0.80$ on gemini ($p = 0.014$), $+0.73$ on haiku ($p = 0.031$) and $+0.67$ on luna ($p = 0.059$), by the same exact permutation test used in §5.2 against the same critical $|\rho| = 0.700$. Two of the three are significant, and the third misses by a hair in the same direction.

What this confounds. The *ordering of techniques* by visibility effect. A reader who takes our per-technique ranking as a statement about which rhetorical strategies engines reward is reading, in substantial part, a statement about how much text each transformation leaves behind. We do not know how to separate the two with this design, because length is not a nuisance variable that could have been held fixed: rewriting a source *is* changing its text, and a rewrite that adds quotations or replaces prose with figures necessarily changes its size. The honest position is that the per-technique effect sizes in Table 4 are joint effects of the rhetorical manipulation and the length change it entails, and we have not decomposed them.

What this does *not* confound. The comparison between engines, which is this paper's claim. The transformation is performed once and the identical text — with the identical length change — is served to all three engines (§3). Any disagreement between them is therefore not attributable to length, because length does not vary between them. Fluency Optimization makes this concrete: it lengthens the source by 7.3%, and that same lengthened text raises visibility significantly on haiku and lowers it significantly on luna. A mechanical length effect cannot produce opposite signs from identical input. Nor does length touch the abstention result of §5.8, which is about whether the engine refuses at all, not about shares.

Why we report it. This analysis was not planned; it was prompted by a reviewer asking whether the metric rewards content rather than credibility, and it partially confirms that worry. We report it because a confound found by the authors and bounded is worth more than one found by a reader after publication, and because the bound turns out to matter: it narrows what our per-technique numbers mean without touching the finding the paper is named for.

5.8 Abstention, and how optimization erodes it

The induction result of §5.6 says that optimized sources get cited on queries they cannot answer. It does not say what the engine does with the rest of the answer. That turns out to be the more serious half.

On the 25 pitfall queries, at baseline, all three engines abstain in 100% of generated answers — they state that the provided sources do not answer the question, usually in a near-verbatim fixed phrase ("As fontes fornecidas não respondem a esta pergunta"). The refusal is not marginal or hedged; it is unanimous, across three model families and every repetition.

technique	engine	n_respostas	n_abstencao	pct_abstencao	queda_vs_baseline_pp
baseline	gemini-3.5-flash-lite	75	75	100.0	n/d
baseline	claude-haiku-4-5	75	75	100.0	n/d
baseline	gpt-5.6-luna	75	75	100.0	n/d
cite_sources	gemini-3.5-flash-lite	75	75	100.0	0
cite_sources	claude-haiku-4-5	75	75	100.0	0
cite_sources	gpt-5.6-luna	75	75	100.0	0
quotation_addition	gemini-3.5-flash-lite	75	70	93.33	6.667
quotation_addition	claude-haiku-4-5	75	72	96.00	4.000
quotation_addition	gpt-5.6-luna	75	72	96.00	4.000
statistics_addition	gemini-3.5-flash-lite	75	51	68.00	32.00
statistics_addition	claude-haiku-4-5	75	60	80.00	20.00
statistics_addition	gpt-5.6-luna	75	56	74.67	25.33
fluency_optimization	gemini-3.5-flash-lite	75	74	98.67	1.333
fluency_optimization	claude-haiku-4-5	75	75	100.0	0
fluency_optimization	gpt-5.6-luna	75	74	98.67	1.333
easy_to_understand	gemini-3.5-flash-lite	75	64	85.33	14.67
easy_to_understand	claude-haiku-4-5	75	72	96.00	4.000
easy_to_understand	gpt-5.6-luna	75	72	96.00	4.000
authoritative	gemini-3.5-flash-lite	75	63	84.00	16.00
authoritative	claude-haiku-4-5	75	66	88.00	12.00
authoritative	gpt-5.6-luna	75	69	92.00	8.000
unique_words	gemini-3.5-flash-lite	75	75	100.0	0
unique_words	claude-haiku-4-5	75	75	100.0	0
unique_words	gpt-5.6-luna	75	75	100.0	0
technical_terms	gemini-3.5-flash-lite	75	71	94.67	5.333
technical_terms	claude-haiku-4-5	75	75	100.0	0
technical_terms	gpt-5.6-luna	75	72	96.00	4.000
keyword_stuffing	gemini-3.5-flash-lite	75	62	82.67	17.33
keyword_stuffing	claude-haiku-4-5	75	67	89.33	10.67
keyword_stuffing	gpt-5.6-luna	75	69	92.00	8.000

Table 12: Abstention on the 25 pitfall queries, by technique and engine: the share of generated answers in which the engine states that the provided sources do not answer the question. At baseline all three engines abstain in 100% of answers; every figure below that is erosion of a correct refusal. Detection is by lexical pattern over PT-BR refusal phrasings, hand-checked on a sample — not a validated classifier (§6); it is reliable here only because the baseline refusal is a near-verbatim fixed phrase.

Optimizing one of the five sources erodes that refusal, and Statistics Addition erodes it most: abstention falls to 68.0% on gemini, 74.7% on luna and 80.0% on haiku — a drop of 32, 25 and 20 percentage points from a unanimous baseline. Keyword Stuffing, Authoritative and Easy-to-Understand follow. Only two techniques — Cite Sources and Unique Words — leave the refusal intact on all three engines.

One of them, Unique Words, is also the second most damaging technique for visibility on answerable queries (-54% and -55% on two engines): a technique can be safe in this sense and useless in the other.

What replaces the refusal is not a hedge. Pitfall query `p006` asks whether a named physiotherapist at a named clinic has an opening next Tuesday — a fact about a specific person's calendar, present in no source. Every baseline answer declines. Under Statistics Addition, the engine answers:

> "A ocupação da agenda do fisioterapeuta André Moura para a próxima terça-feira oscila constantemente [...] havendo apenas de 3% a 5% de vagas remanescentes disponíveis [4]."
>
> *("The occupancy of physiotherapist André Moura's calendar for next Tuesday fluctuates constantly [...] with only 3% to 5% of remaining slots available [4].")*

The number is invented, the question is unanswerable, and the claim is attributed to source [4]. The technique did not merely raise a source's share of the answer: it converted a correct refusal into a fabricated, sourced assertion.

This failure transfers where the benefit does not. The comparison with §5.2 is the point of this subsection. Statistics Addition's effect on *visibility* is as engine-dependent as anything in this study — indeterminate on gemini (-1.4%) and haiku ($+0.6\%$), catastrophic on luna (-81.5%). Its effect on *abstention* is consistent: all three engines lose 20 to 32 points. A practitioner choosing techniques by visibility would find this one unusable on one engine and neutral on two; the safety cost is paid on all three regardless. Portability, in this benchmark, is a property of the damage rather than of the gain — and that holds not only across techniques, as §5.2 showed, but across the two kinds of outcome.

We are careful about what this measurement is. Abstention is detected by a lexical pattern over PT-BR refusal phrasings, hand-checked on a sample, not by a validated classifier or human annotation; it works here because the baseline refusal is so stereotyped. The direction and the size of the drop are too large to be an artifact of phrasing, but the exact percentages do not carry the guarantee that the visibility measurements do, which come from word counts. §6 records what a proper validation would require.

5.9 Cost

categoria	n_chamadas_engine	tokens_input_medio	tokens_output_medio	custo_usd_total_engine	custo_usd_medio_chamada	n_chamadas_transform	tokens_input_medio_transform	tokens_output_medio_transform	custo_usd_total_transform	custo_usd_total
baseline	1575	2713.1	312.0	1.275	0.0008	0	n/d	n/d	0	1.275
cite_sources	1575	2727.4	312.0	1.279	0.0008	525	636.8	490.8	0.3788	1.657
quotation_addition	1575	2712.2	309.6	1.269	0.0008	525	626.8	475.5	0.3682	1.638
statistics_addition	1575	2714.0	314.1	1.280	0.0008	525	637.8	477.4	0.3704	1.651
fluency_optimization	1575	2728.8	311.0	1.277	0.0008	525	629.8	492.1	0.3793	1.657
easy_to_understand	1575	2595.4	311.2	1.245	0.0008	525	633.8	358.8	0.2907	1.536
authoritative	1575	2654.9	309.1	1.255	0.0008	525	630.8	418.2	0.3303	1.586
unique_words	1575	2751.4	313.9	1.289	0.0008	525	628.8	515.2	0.3944	1.683
technical_terms	1575	2758.3	317.6	1.297	0.0008	525	621.8	521.7	0.3983	1.695
keyword_stuffing	1575	2629.9	304.6	1.241	0.0008	525	630.8	393.3	0.3136	1.555
TOTAL	15750	n/d	n/d	n/d	n/d	4725	n/d	n/d	n/d	15.93

Table 13: API cost table: tokens and US$ per technique, aggregated directly from execution traces.

The table above covers the primary engine, whose full run — $15{,}750$ engine calls plus $4{,}725$ transformation calls — cost US$15.93 under batch pricing. Cost per technique is nearly flat, from US$1.54 (Easy-to-Understand) to US$1.70 (Technical Terms), a spread of about 10% that reflects output length rather than anything about the technique's effectiveness: the cheapest and the most expensive techniques are both penalties on this engine.

The three-engine study totals US$66.69, and the distribution is uneven in a way worth recording for anyone planning a replication: the same 525-query benchmark cost US$15.93, US$44.66 and US$6.10 on `gemini-3.5-flash-lite`, `claude-haiku-4-5` and `gpt-5.6-luna` respectively — a factor of more than seven between the cheapest and the most expensive, at the same tier of model and the same workload. Per-call cost is logged in every trace and summed from those records rather than recomputed from a price table (§4), so these are amounts actually spent.

6 Limitations

We report these limitations as integral to interpreting the results above, not as an afterthought.

1. **Fixed context; no measurement of discoverability.** This study's protocol places all five candidate sources in the engine's context *before* measuring anything, exactly as in Aggarwal et al. [2024]. It therefore measures the effect of a GEO technique on the citation/summarization step, *conditioned on the source already being retrieved* — it does not, and cannot, measure whether an optimized page would be more (or less) likely to be crawled, indexed, or retrieved by a real generative engine's search stage in the first place, nor any durable effect on real-world traffic. Any claim about "visibility" in this paper should be read narrowly, as citation-share-given-retrieval, not as end-to-end search visibility.

   This is not a peripheral caveat, and the field has named it. Surveying 45 GEO studies, Martinez [2026] models generative search as a multi-stage pipeline and argues that results obtained with the source "already present in a five-document context" establish "neither organic discoverability nor durable traffic effects." The survey further reports that content-side rewriting can move the two stages in *opposite* directions — reducing a document's presence in the retrieved set while improving how it is treated once retrieved, which it describes as "a downstream gain masking an upstream loss." Our design observes only the downstream stage by construction. A technique we measure as harmless, or even mildly positive, could still be a net loss end to end, and nothing in this paper would reveal it. We regard closing that gap — measuring the retrieval stage under the same controlled conditions — as the more consequential extension of this work than adding further engines.

2. **Synthetic-realistic, not scraped, sources.** The five sources per query are generated to be realistic PT-BR web content (by a model distinct from either evaluation engine, see §4) rather than scraped from real, live web pages. This was a deliberate scope decision (recorded in this project's `PROGRESS.md`) accepted for the pilot and, at the time of writing, for the full run; each source's origin is labeled in the dataset's `origem` field for full transparency. Findings about which techniques help or hurt visibility should be understood as findings about text *as measured by this study's synthetic corpus*, not as a claim that has been validated against the noisier structure (formatting, metadata, boilerplate, mixed quality) of real scraped PT-BR web pages.

3. **The transformation side is never cross-engine.** All three engines answer over the *same* already-transformed source texts, produced once by `gemini-3.5-flash-lite` (§4). That is the right design for the question we ask — it isolates citation behavior from transformation quality, and without it a cross-engine difference could not be attributed to the engine at all — but it bounds the claim: this study measures whether the *citation-side* effect transfers across engines, never whether a different engine would have produced a different, and possibly more effective, transformation of the same source in the first place. A technique that underperforms on one engine here may be underperforming because of how a Gemini-family model chose to rewrite the text, not because that engine is inherently unresponsive to the technique. Re-running the transformation step per engine would test that, at three times the transformation cost, and is the most direct extension of this work.

4. **Engines were not collected in one shared time window, and one does not accept a custom temperature.** Two asymmetries across the three engines are forced rather than chosen. First, temperature: the GPT-5 family does not accept a custom temperature, so `gpt-5.6-luna` runs at its provider default while the other two run at 0.7; a difference in sampling temperature is a plausible contributor to any difference in how sharply an engine reacts, and cannot be separated from the engine identity in this design. Second, timing: the study was extended from one engine to three after the first was already complete, so while every trace of a given engine carries a single `model_version` (the guarantee the protocol requires against mid-collection model updates), the three runs are spread over days rather than sharing one window. The paired design protects the comparison from query-composition differences, but not from a provider silently changing serving behavior behind a stable version string.

5. **Self-preference risk on the transformation side.** The nine GEO techniques are applied to source text by the *same* model family that also serves as the primary evaluation engine (`gemini-3.5-flash-lite` transforms sources; a Gemini-family model also judges/cites them in the baseline and technique conditions) — matching the original paper's own choice to use `gpt3.5-turbo` for both roles, but

carrying the same risk: an engine could systematically favor content it (or a closely related model) generated, inflating apparent technique effectiveness. Two design choices bound, but do not eliminate, this risk: (i) queries and source *content* originate from a third model family (Grok), unrelated to either judging engine, so the underlying material being cited is not self-generated by the judge; and (ii) two of the three evaluation engines (`claude-haiku-4-5`, `gpt-5.6-luna`) are unrelated to the transforming model family, so §5.2 directly observes how non-Gemini judges treat Gemini-transformed text. Where a technique helps on Gemini and hurts on both others, self-preference on the transformation side is a candidate explanation that this design cannot rule out and does not claim to; that ambiguity is precisely why the transformation step would have to be re-run per engine to settle it.

6. **Batch collection can fail non-randomly, and did.** Batch APIs report per-job success, not per-cell success, and a job can be marked collected while a contiguous block of its cells failed — in our case when an account credit balance was exhausted mid-job. On one engine this left a full complement of output files while 41 of 525 queries were silently incomplete, and the missing queries were *not* missing at random: they formed one contiguous block of query identifiers, because the failure was a single budget event rather than scattered timeouts. The consequence is not hypothetical: at the incomplete n, one technique's uncorrected confidence interval crossed zero, and it excluded zero on both affected engines once the missing cells were recovered — the same two arms that Holm later attenuates (§5.2); the point is how far the missing block moved the estimate, not the arms' final status. We therefore treat completeness as a *per-cell* property — baseline plus all nine techniques, each with all repetitions present — verified before any analysis, and never infer completeness from output-file counts. We report this as a limitation of batch-collected LLM evaluation generally, not only of this run: any study that measures progress by files rather than by cells can carry a systematically absent block of data that is large enough to invert a reported finding.

7. **PT-BR only.** All queries, sources, and engine interactions are in Brazilian Portuguese. Findings should not be assumed to transfer to other languages or markets (including European Portuguese) without separate validation; PT-BR was chosen for its direct relevance to the Brazilian market this project targets, not as a stand-in for Portuguese-language behavior in general.

8. **Absolute visibility scores are not comparable across studies.** Imp_{wc} and Imp_{pwc} in this paper are on a $[0, 1]$ share-of-answer scale, by construction; the original paper reports visibility on its own per-study normalized scale. Every comparison against Aggarwal et al. [2024] in this paper (§5.5) is restricted to sign and rank correlation; treating an absolute GEO-PTBR number as comparable in magnitude to an absolute number from the original paper's tables would be a methodological error, and this paper does not do so anywhere.

9. **Fewer repetitions than the original study.** Each cell is repeated 3 times (temperature 0.7) versus the original paper's 5 random seeds, a cost-driven deviation (§3.1) that yields wider confidence intervals for a fixed query budget than the original design would.

10. **Forced engine substitutions, not the originally specified model.** The primary engine differs from what this study's original contract specified (`gemini-2.5-flash` → `gemini-3.6-flash` → `gemini-3.5-flash-lite`), driven by API account availability and cost, not by scientific preference — recorded in full in this project's `TASK.md` §E6. The move from a single-engine design with a 50-query cross-validation sample to three full-benchmark engines was likewise a mid-study change, made once the dataset and harness costs were already sunk; the three engines are matched by provider tier, which is the best available proxy for comparability but is not a controlled matching on capability. In particular, `gemini-3.5-flash-lite` is a smaller, cheaper model than the larger `-flash` tier originally planned, and may follow the (sometimes elaborate) transformation instructions with less fidelity than a larger model would; this is a plausible confound for any technique whose GEO instruction is more complex to execute correctly (e.g., Statistics Addition, which requires generating

plausible, internally consistent numbers) relative to techniques that require only a stylistic pass (e.g., Fluency Optimization).

11. **The two-stage decomposition this dataset supports, and that we do not fit.** A percentage change in citation share conflates two things: whether the source is cited at all, and how much of the answer it gets once cited. The zeroing rates we report make the first stage visible — Unique Words removes the source entirely from 26.1% of answers on haiku and 20.6% on luna, against 0.0% on gemini — but the headline estimand remains the pooled one, for comparability with the original paper's Eq. 4. A cleaner analysis would model $P(\text{cited}) \times E[\text{share} \mid \text{cited}]$ as two endpoints. The released traces contain everything needed to do it, and we flag it as the most obvious analysis this dataset supports and this paper does not perform.

12. **No manipulation check on the transformations.** We assume each transformed source instantiates its intended technique and preserves the original's factual content. Neither is audited. No human or independent model verified that the Statistics Addition output actually adds statistics, that Authoritative actually reads as more authoritative, or that any transformation left the source's claims intact — and §5.8 shows transformations can manufacture answer-bearing content that was not there. We measured the one component of this that is mechanically measurable, the change in source length (§5.7), and found it rank-correlated with the visibility effect on all three engines. The rest — fidelity to the technique, factual preservation, answer-equivalence to the original — remains unaudited, and until it is, per-technique effects cannot be attributed cleanly to style, authority cues or terminology rather than to changed content. An independent audit of a sample of transformed sources is the cheapest high-value addition to this study.

13. **The metric rewards content, and some techniques add content.** Visibility is a share of the generated answer attributable to a source, so a transformation that makes a source carry more usable material can raise it without the engine having judged the source more credible. This is sharpest for Statistics Addition, Quotation Addition and Cite Sources, whose instructions add quantitative claims, attributed quotes and references — these are *content* transformations, not purely stylistic ones, and in the limit a transformed source can answer something the original did not. Our pitfall control detects the extreme form of this (§5.6), but the metric cannot separate "cited more because it reads as more authoritative" from "cited more because it now contains more of the answer."

14. **The pitfall sources are unanswerable by construction, not by independent adjudication.** The 25 control queries were designed so that no provided source answers them, and the design is verifiable from the data — each carries an explicit marker, and baseline visibility on them is exactly zero on all three engines, which is three independent judges behaving as if the sources do not answer. That is convergent evidence, not validation: we did not run human annotators, blind to condition, rating whether each source suffices to answer, with an inter-rater statistic. For a finding we ourselves frame as a safety result, that check is cheap and should exist; it does not yet.

15. **Everything here is measured at the economy tier, and stops there.** All three engines are their providers' small/economy models. That was a budget decision (§5.9), and it bounds the claim: what this study demonstrates is that the effect of a fixed GEO technique is engine-dependent *among economy-tier engines*, on three of them, in PT-BR. Whether frontier-tier models exhibit the same instability is not measured here and should not be inferred from these results in either direction. Two opposite hypotheses are equally consistent with what we report and equally untested by it: that larger models follow citation behavior more stably, so the disagreement we measure would shrink; or that the disagreement reflects genuinely different learned preferences over source text, in which case tier would not remove it. Settling this needs a frontier-tier run, which is a straightforward extension of the released harness — the engine is a configuration value — and the reason we did not do it is cost, not design.

7 Conclusion

GEO-PTBR provides the first Brazilian-Portuguese replication of the GEO protocol [Aggarwal et al., 2024], executed end-to-end against a live generative-engine API with per-trace cost and provenance logging, and executed in full against three generative engines from three different model families.

Our headline result is not a technique but a boundary condition: **the effect of a GEO technique depends on which engine answers**. Over identical queries and identical optimized sources, the three engines agree on the direction of the effect for only four of nine techniques, and all four are techniques that reduce visibility. Fluency Optimization — the mildest transformation in the catalogue, which only rewrites the source to read better — inverts outright: significantly positive on one engine and significantly negative on another, both arms surviving Holm correction over the primary family. The larger-looking inversion of Technical Terms does not survive that correction, and we report it as the attenuation it becomes rather than the inversion it first appeared to be. Agreement itself falls from 66.7% with two engines to 44.4% with three, so a two-engine cross-validation would have overstated how far these results travel.

Against the original English study, five of nine effect directions match, but the level does not: where the original reports its best techniques at $+27$ to $+41\%$, our largest positive effect on any engine is $+2.6\%$, and the largest effect anywhere in the study is a penalty of -81.5%. We do not claim the technique *ordering* transfers — under the primary (median) estimator, an exact permutation test leaves the rank correlations indistinguishable from chance, and at $n = 9$ this study is underpowered to settle ordering either way (§5.5). Sector behavior compounds this — health, the YMYL sector, is the least responsive of the three, and the only effect stable across all sectors is the harm done by Keyword Stuffing.

We report what did not work with the same weight as what did, because in this benchmark it is most of the result. Six of nine techniques are either indistinguishable from no effect or actively harmful on the primary engine, and five of nine induce citation of sources that cannot answer the query at all — an outcome that is a safety failure rather than an optimization, and one that a study measuring only visibility on answerable queries would not have seen. It is worth naming where that failure lands: nine of the twenty-five pitfall queries are health questions, so the techniques that most reliably induce a citation of a source that cannot answer are inducing it, in a third of cases, on YMYL content.

Stated at the precision the design actually supports, the boundary condition reads: the effect of a GEO technique *as executed by one transforming model* depends on which engine answers. The transformation step is held fixed by construction, which is what makes the comparison internally valid — all three engines read the identical text — and is also what the result is conditional on. Whether a technique executed by a different transforming model would produce the same disagreement is a separate experiment, and the released harness runs it by changing one configuration value.

A third boundary is internal to the numbers themselves. How much a technique changes the source's length rank-predicts how much it changes the source's visibility ($\rho = +0.80$, $+0.73$, $+0.67$ across the three engines), so our ordering of techniques is partly an ordering by a mechanical property of the rewrite. It does not reach the engine comparison — identical text, identical length change, opposite signs — but it does narrow what a per-technique number means.

Two further boundaries keep that result from being read as more than it is. All three engines are economy-tier, so what we demonstrate is engine-dependence *within* that tier; whether frontier models are steadier is untested here, and assuming they are is as unsupported as assuming they are not. And the sources are optimized text, not live web pages. Neither boundary rescues the fixed-catalogue approach — a technique catalogue that does not survive three engines of one tier in one language has not earned the word "portable" — but both mark where the evidence stops.

These results are the experimental counterpart of a claim the adaptive-GEO literature makes from the other direction. Wu et al. [2026] and Yuan et al. [2026] both begin from the premise that a fixed catalogue of hand-designed transformations is the wrong primitive — the first learning engine preference rules, the second evolving compositional strategies because static heuristics "cannot flexibly adapt to [...] the changing behaviors of generative engines." That premise is normally argued and then assumed.

We measured it: applied unchanged to identical sources and queries, the nine transformations agree in direction on four of nine across three engines, all four of them harmful, and the mildest of them inverts sign between two engines with both arms surviving correction. If a fixed catalogue carried a stable effect from one engine to the next, the case for learning or evolving engine-specific strategies would be weaker; our measurement is that it does not. We make no claim about how well those adaptive methods themselves transfer — we did not evaluate them — and Martinez [2026] reaches the same negative conclusion from the literature that we reach from a controlled run.

One scope note belongs with that. Our three sectors — health, legal, real estate — were chosen because two are YMYL and one is high-value transactional, which is where citation errors are costly and where the regulatory attention is. It is a deliberate choice and not a sample of domains: e-commerce, consumer technology and general informational content are absent, and nothing here establishes that the anti-hack pattern we observe holds outside the three we ran.

What this leaves a practitioner. A fair objection to a study whose headline is instability: if the effect depends on the engine and we tested one tier of three engines, what is anyone supposed to do with it? The honest answer is short, and it is a result rather than a shrug. First: do not port an optimization from one engine to another without measuring it on the second, because in this benchmark five of nine techniques changed sign or lost their effect across engines, and the mildest technique in the catalogue inverted. Second: the techniques that transferred are the ones that *hurt*, so the safest use of a fixed catalogue is as a list of things to avoid rather than to apply — Keyword Stuffing costs visibility on all three engines, with no engine on which it pays. Third, and least comfortable: some techniques that look like optimization are also reliability failures. Statistics Addition erodes a correct refusal on every engine we tested (§5.8), and that cost does not depend on which engine answers, so it is not something a per-engine test would reveal. None of this requires knowing which technique wins; all of it follows from measurement rather than from the catalogue.

The practical implication is that "optimizing for AI search" is not a single, portable target: a content strategy validated on one generative engine may be neutral on a second and actively counterproductive on a third. The lasting contribution we intend is therefore not any single number above but the released artifact — 525 PT-BR queries with sources, per-technique results on three engines, and the full measurement pipeline — so that the boundary we found can be tested, contested, and extended on engines and languages we did not cover.

A Complete tables

Every table in the body is a curated view: the full CSV carries auxiliary columns (mean-based estimates, level confidence intervals, zero and undefined counts) and, for some tables, rows for the secondary metric and query set. Those views were chosen for legibility, not to hide anything. This appendix removes the need to take that on trust: below is every table on which a claim in this paper rests, with all columns and all rows, from the same CSVs that produced the body versions. Nothing is recomputed — the two are emitted by the same script from the same file, differing only in which columns are printed. Where a table is too wide for the page it is split by columns into parts, with the identifying columns repeated in each part, so any row can still be read end to end.

Two tables are deliberately *not* reproduced here in full: the breakdowns by sector and by target-source position. They are the two analyses this paper declares exploratory rather than confirmatory (§5.2), they are the two largest (108 and 180 rows, which column-splitting would multiply into some sixty pages of six-point type), and their body views already show the primary query set and metric — the rows any claim would use. Both are released complete in `results/`, with the study version and model identifiers in their header lines, as is every table below.

set	technique	metric	n	n_baseline_zero	n_indefinido	base.	baseline_ci_lo	baseline_ci_hi	tech.	tecnica_ci_lo	tecnica_ci_hi
baseline_pos	cite_sources	pwc	497	0	0	0.1474	0.1434	0.1513	0.1511	0.1474	0.1547
all	cite_sources	pwc	500	3	2	0.1465	0.1425	0.1506	0.1505	0.1469	0.1543
baseline_pos	cite_sources	wc	497	0	0	0.1998	0.1949	0.2046	0.2056	0.2010	0.2100
all	cite_sources	wc	500	3	2	0.1986	0.1936	0.2037	0.2048	0.2002	0.2096
baseline_pos	quotation_addition	pwc	497	0	0	0.1474	0.1434	0.1513	0.1481	0.1446	0.1516
all	quotation_addition	pwc	500	3	2	0.1465	0.1425	0.1506	0.1473	0.1437	0.1510
baseline_pos	quotation_addition	wc	497	0	0	0.1998	0.1949	0.2046	0.2021	0.1978	0.2063
all	quotation_addition	wc	500	3	2	0.1986	0.1936	0.2037	0.2011	0.1967	0.2056
baseline_pos	statistics_addition	pwc	497	0	0	0.1474	0.1434	0.1513	0.1471	0.1430	0.1511
all	statistics_addition	pwc	500	3	1	0.1465	0.1425	0.1506	0.1464	0.1422	0.1506
baseline_pos	statistics_addition	wc	497	0	0	0.1998	0.1949	0.2046	0.2017	0.1964	0.2069
all	statistics_addition	wc	500	3	1	0.1986	0.1936	0.2037	0.2007	0.1954	0.2061
baseline_pos	fluency_optimization	pwc	497	0	0	0.1474	0.1434	0.1513	0.1501	0.1463	0.1537
all	fluency_optimization	pwc	500	3	1	0.1465	0.1425	0.1506	0.1493	0.1455	0.1531
baseline_pos	fluency_optimization	wc	497	0	0	0.1998	0.1949	0.2046	0.2055	0.2009	0.2099
all	fluency_optimization	wc	500	3	1	0.1986	0.1936	0.2037	0.2044	0.1998	0.2090
baseline_pos	easy_to_understand	pwc	497	0	0	0.1474	0.1434	0.1513	0.1379	0.1339	0.1419
all	easy_to_understand	pwc	500	3	2	0.1465	0.1425	0.1506	0.1376	0.1335	0.1416
baseline_pos	easy_to_understand	wc	497	0	0	0.1998	0.1949	0.2046	0.1892	0.1843	0.1940
all	easy_to_understand	wc	500	3	2	0.1986	0.1936	0.2037	0.1887	0.1836	0.1937
baseline_pos	authoritative	pwc	497	0	0	0.1474	0.1434	0.1513	0.1434	0.1394	0.1472
all	authoritative	pwc	500	3	1	0.1465	0.1425	0.1506	0.1427	0.1388	0.1468
baseline_pos	authoritative	wc	497	0	0	0.1998	0.1949	0.2046	0.1945	0.1898	0.1992
all	authoritative	wc	500	3	1	0.1986	0.1936	0.2037	0.1936	0.1888	0.1985
baseline_pos	unique_words	pwc	497	0	0	0.1474	0.1434	0.1513	0.1424	0.1385	0.1460
all	unique_words	pwc	500	3	1	0.1465	0.1425	0.1506	0.1416	0.1377	0.1454
baseline_pos	unique_words	wc	497	0	0	0.1998	0.1949	0.2046	0.1927	0.1881	0.1971
all	unique_words	wc	500	3	1	0.1986	0.1936	0.2037	0.1917	0.1869	0.1964
baseline_pos	technical_terms	pwc	497	0	0	0.1474	0.1434	0.1513	0.1522	0.1484	0.1560
all	technical_terms	pwc	500	3	2	0.1465	0.1425	0.1506	0.1514	0.1477	0.1553
baseline_pos	technical_terms	wc	497	0	0	0.1998	0.1949	0.2046	0.2073	0.2027	0.2119
all	technical_terms	wc	500	3	2	0.1986	0.1936	0.2037	0.2062	0.2016	0.2111
baseline_pos	keyword_stuffing	pwc	497	0	0	0.1474	0.1434	0.1513	0.1314	0.1270	0.1358
all	keyword_stuffing	pwc	500	3	1	0.1465	0.1425	0.1506	0.1307	0.1263	0.1350
baseline_pos	keyword_stuffing	wc	497	0	0	0.1998	0.1949	0.2046	0.1751	0.1698	0.1803
all	keyword_stuffing	wc	500	3	1	0.1986	0.1936	0.2037	0.1741	0.1687	0.1794

Table 14: Main results: position-adjusted word count (Imp_{pwc}) and word count (Imp_{wc}) of the target source, baseline vs. technique, with bootstrap 95% CI and relative improvement (mean and median, Eq. 4). Each technique appears twice, once per query set: `baseline_pos` (**primary**) restricts to queries whose target source has positive baseline visibility, and `all` (sensitivity) adds the queries where the source was never cited, which enter Eq. 4 as an exact 0% and pull the median toward zero. Read the primary row as the result; see §3.5. **Part 1 of 4** of the complete table; the identifying columns repeat in every part.

set	technique	metric	n	mean %	melhoria_media_ci_lo	melhoria_media_ci_hi	median %	CI lo	CI hi
baseline_pos	cite_sources	pwc	497	15.35	7.096	27.07	1.809	0.2001	3.904
all	cite_sources	pwc	500	15.32	7.157	26.78	1.740	0.2001	3.796
baseline_pos	cite_sources	wc	497	10.76	5.998	16.70	2.218	0.3898	3.657
all	cite_sources	wc	500	10.73	5.982	16.68	2.175	0.3590	3.657
baseline_pos	quotation_addition	pwc	497	11.77	3.378	26.04	0.5429	-0.6744	2.595
all	quotation_addition	pwc	500	11.75	3.428	26.04	0.5363	-0.6521	2.132
baseline_pos	quotation_addition	wc	497	7.763	3.098	14.48	0.3026	-0.0803	1.558
all	quotation_addition	wc	500	7.747	3.145	14.40	0.2370	-0.0635	1.558
baseline_pos	statistics_addition	pwc	497	11.02	4.707	18.26	-1.347	-3.286	0.7224
all	statistics_addition	pwc	500	10.98	4.761	18.28	-1.324	-3.114	0.5866
baseline_pos	statistics_addition	wc	497	8.769	4.071	14.16	0	-2.153	1.436
all	statistics_addition	wc	500	8.734	4.012	14.11	0	-2.104	1.275
baseline_pos	fluency_optimization	pwc	497	10.75	5.579	17.36	0.6723	-0.3636	2.483
all	fluency_optimization	pwc	500	10.71	5.459	17.26	0.6499	-0.3601	2.313
baseline_pos	fluency_optimization	wc	497	9.719	5.578	14.71	0.7321	0	2.292
all	fluency_optimization	wc	500	9.680	5.432	14.70	0.5522	0	2.201
baseline_pos	easy_to_understand	pwc	497	5.697	-3.958	20.87	-5.365	-7.573	-3.571
all	easy_to_understand	pwc	500	5.686	-3.885	20.33	-5.255	-7.573	-3.459
baseline_pos	easy_to_understand	wc	497	1.506	-3.765	8.621	-3.070	-4.589	-1.497
all	easy_to_understand	wc	500	1.503	-3.818	8.502	-3.068	-4.553	-1.383
baseline_pos	authoritative	pwc	497	8.970	0.5160	22.15	-2.358	-4.307	-0.2585
all	authoritative	pwc	500	8.934	0.5457	22.03	-2.343	-4.193	0
baseline_pos	authoritative	wc	497	4.661	-0.5040	11.56	-2.175	-3.733	0
all	authoritative	wc	500	4.642	-0.4681	11.36	-2.120	-3.733	0
baseline_pos	unique_words	pwc	497	1.817	-1.698	5.748	-2.139	-4.166	-0.2178
all	unique_words	pwc	500	1.810	-1.731	5.695	-2.062	-4.230	-0.2078
baseline_pos	unique_words	wc	497	0.3775	-2.567	3.805	-1.691	-3.264	-0.5802
all	unique_words	wc	500	0.3760	-2.607	3.681	-1.665	-3.249	-0.5139
baseline_pos	technical_terms	pwc	497	11.73	7.334	16.75	2.553	0.9487	4.282
all	technical_terms	pwc	500	11.70	7.351	16.68	2.526	0.9487	4.084
baseline_pos	technical_terms	wc	497	9.871	6.395	13.86	2.546	1.152	4.109
all	technical_terms	wc	500	9.851	6.344	13.81	2.534	1.072	3.969
baseline_pos	keyword_stuffing	pwc	497	-4.019	-9.032	1.955	-9.843	-11.70	-5.697
all	keyword_stuffing	pwc	500	-4.003	-9.010	1.992	-9.751	-11.66	-5.590
baseline_pos	keyword_stuffing	wc	497	-7.724	-11.44	-3.755	-8.569	-11.30	-6.442
all	keyword_stuffing	wc	500	-7.693	-11.37	-3.590	-8.423	-10.93	-6.409

Table 15: Main results: position-adjusted word count (Imp_{pwc}) and word count (Imp_{wc}) of the target source, baseline vs. technique, with bootstrap 95% CI and relative improvement (mean and median, Eq. 4). Each technique appears twice, once per query set: `baseline_pos` (**primary**) restricts to queries whose target source has positive baseline visibility, and `all` (sensitivity) adds the queries where the source was never cited, which enter Eq. 4 as an exact 0% and pull the median toward zero. Read the primary row as the result; see §3.5. **Part 2 of 4** of the complete table; the identifying columns repeat in every part.

set	technique	metric	n	CI excl. 0 (mean)	CI excl. 0	p	p (Holm)	sig. (Holm)	p (BH)
baseline_pos	cite_sources	pwc	497	yes (+)	yes (+)	0.0252	0.1512	no	0.0513
all	cite_sources	pwc	500	yes (+)	yes (+)	0.0256	n/d	sens.	n/d
baseline_pos	cite_sources	wc	497	yes (+)	yes (+)	0.0030	n/d	sens.	n/d
all	cite_sources	wc	500	yes (+)	yes (+)	0.0032	n/d	sens.	n/d
baseline_pos	quotation_addition	pwc	497	yes (+)	no	0.3472	0.5364	no	0.3472
all	quotation_addition	pwc	500	yes (+)	no	0.3550	n/d	sens.	n/d
baseline_pos	quotation_addition	wc	497	yes (+)	no	0.4972	n/d	sens.	n/d
all	quotation_addition	wc	500	yes (+)	no	0.5146	n/d	sens.	n/d
baseline_pos	statistics_addition	pwc	497	yes (+)	no	0.1788	0.5364	no	0.2299
all	statistics_addition	pwc	500	yes (+)	no	0.2106	n/d	sens.	n/d
baseline_pos	statistics_addition	wc	497	yes (+)	no	0.9538	n/d	sens.	n/d
all	statistics_addition	wc	500	yes (+)	no	1.000	n/d	sens.	n/d
baseline_pos	fluency_optimization	pwc	497	yes (+)	no	0.2672	0.5364	no	0.3006
all	fluency_optimization	pwc	500	yes (+)	no	0.2940	n/d	sens.	n/d
baseline_pos	fluency_optimization	wc	497	yes (+)	no	0.1608	n/d	sens.	n/d
all	fluency_optimization	wc	500	yes (+)	no	0.1850	n/d	sens.	n/d
baseline_pos	easy_to_understand	pwc	497	no	yes (−)	0.0001	0.0009	yes (−)	0.0004
all	easy_to_understand	pwc	500	no	yes (−)	0.0001	n/d	sens.	n/d
baseline_pos	easy_to_understand	wc	497	no	yes (−)	0.0001	n/d	sens.	n/d
all	easy_to_understand	wc	500	no	yes (−)	0.0001	n/d	sens.	n/d
baseline_pos	authoritative	pwc	497	yes (+)	yes (−)	0.0342	0.1520	no	0.0513
all	authoritative	pwc	500	yes (+)	yes (−)	0.0404	n/d	sens.	n/d
baseline_pos	authoritative	wc	497	no	yes (−)	0.0398	n/d	sens.	n/d
all	authoritative	wc	500	no	no	0.0518	n/d	sens.	n/d
baseline_pos	unique_words	pwc	497	no	yes (−)	0.0304	0.1520	no	0.0513
all	unique_words	pwc	500	no	yes (−)	0.0330	n/d	sens.	n/d
baseline_pos	unique_words	wc	497	no	yes (−)	0.0030	n/d	sens.	n/d
all	unique_words	wc	500	no	yes (−)	0.0026	n/d	sens.	n/d
baseline_pos	technical_terms	pwc	497	yes (+)	yes (+)	0.0032	0.0224	yes (+)	0.0096
all	technical_terms	pwc	500	yes (+)	yes (+)	0.0020	n/d	sens.	n/d
baseline_pos	technical_terms	wc	497	yes (+)	yes (+)	0.0016	n/d	sens.	n/d
all	technical_terms	wc	500	yes (+)	yes (+)	0.0016	n/d	sens.	n/d
baseline_pos	keyword_stuffing	pwc	497	no	yes (−)	0.0001	0.0009	yes (−)	0.0004
all	keyword_stuffing	pwc	500	no	yes (−)	0.0001	n/d	sens.	n/d
baseline_pos	keyword_stuffing	wc	497	yes (−)	yes (−)	0.0001	n/d	sens.	n/d
all	keyword_stuffing	wc	500	yes (−)	yes (−)	0.0001	n/d	sens.	n/d

Table 16: Main results: position-adjusted word count (Imp_{pwc}) and word count (Imp_{wc}) of the target source, baseline vs. technique, with bootstrap 95% CI and relative improvement (mean and median, Eq. 4). Each technique appears twice, once per query set: `baseline_pos` (**primary**) restricts to queries whose target source has positive baseline visibility, and `all` (sensitivity) adds the queries where the source was never cited, which enter Eq. 4 as an exact 0% and pull the median toward zero. Read the primary row as the result; see §3.5. **Part 3 of 4** of the complete table; the identifying columns repeat in every part.

set	technique	metric	n	sig. (BH)
baseline_pos	cite_sources	pwc	497	não (BH)
all	cite_sources	pwc	500	sens.
baseline_pos	cite_sources	wc	497	sens.
all	cite_sources	wc	500	sens.
baseline_pos	quotation_addition	pwc	497	não (BH)
all	quotation_addition	pwc	500	sens.
baseline_pos	quotation_addition	wc	497	sens.
all	quotation_addition	wc	500	sens.
baseline_pos	statistics_addition	pwc	497	não (BH)
all	statistics_addition	pwc	500	sens.
baseline_pos	statistics_addition	wc	497	sens.
all	statistics_addition	wc	500	sens.
baseline_pos	fluency_optimization	pwc	497	não (BH)
all	fluency_optimization	pwc	500	sens.
baseline_pos	fluency_optimization	wc	497	sens.
all	fluency_optimization	wc	500	sens.
baseline_pos	easy_to_understand	pwc	497	yes (−)
all	easy_to_understand	pwc	500	sens.
baseline_pos	easy_to_understand	wc	497	sens.
all	easy_to_understand	wc	500	sens.
baseline_pos	authoritative	pwc	497	não (BH)
all	authoritative	pwc	500	sens.
baseline_pos	authoritative	wc	497	sens.
all	authoritative	wc	500	sens.
baseline_pos	unique_words	pwc	497	não (BH)
all	unique_words	pwc	500	sens.
baseline_pos	unique_words	wc	497	sens.
all	unique_words	wc	500	sens.
baseline_pos	technical_terms	pwc	497	yes (+)
all	technical_terms	pwc	500	sens.
baseline_pos	technical_terms	wc	497	sens.
all	technical_terms	wc	500	sens.
baseline_pos	keyword_stuffing	pwc	497	yes (−)
all	keyword_stuffing	pwc	500	sens.
baseline_pos	keyword_stuffing	wc	497	sens.
all	keyword_stuffing	wc	500	sens.

Table 17: Main results: position-adjusted word count (Imp_{pwc}) and word count (Imp_{wc}) of the target

Table 18: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Part 1 of 5** of the complete table; the identifying columns repeat in every part.

set	metric	technique	engine	n	n_baseline_zero	n_indefinido	base.	baseline_ci_lo	baseline_ci_hi	tech.	tecnica_ci_lo
paired	pwc	cite_sources	gemini-3.5-flash-lite	500	3	2	0.1465	0.1425	0.1506	0.1505	0.1469
paired	pwc	cite_sources	claude-haiku-4-5	500	67	32	0.1337	0.1244	0.1434	0.1376	0.1280
paired	pwc	cite_sources	gpt-5.6-luna	500	62	9	0.0893	0.0840	0.0947	0.0773	0.0723
paired	pwc	quotation_addition	gemini-3.5-flash-lite	500	3	2	0.1465	0.1425	0.1506	0.1473	0.1437
paired	pwc	quotation_addition	claude-haiku-4-5	500	67	31	0.1337	0.1244	0.1434	0.1353	0.1258
paired	pwc	quotation_addition	gpt-5.6-luna	500	62	9	0.0893	0.0840	0.0947	0.0808	0.0757
paired	pwc	statistics_addition	gemini-3.5-flash-lite	500	3	1	0.1465	0.1425	0.1506	0.1464	0.1422
paired	pwc	statistics_addition	claude-haiku-4-5	500	67	33	0.1337	0.1244	0.1434	0.1393	0.1294
paired	pwc	statistics_addition	gpt-5.6-luna	500	62	3	0.0893	0.0840	0.0947	0.0339	0.0296
paired	pwc	fluency_optimization	gemini-3.5-flash-lite	500	3	1	0.1465	0.1425	0.1506	0.1493	0.1455
paired	pwc	fluency_optimization	claude-haiku-4-5	500	67	46	0.1337	0.1244	0.1434	0.1522	0.1424
paired	pwc	fluency_optimization	gpt-5.6-luna	500	62	13	0.0893	0.0840	0.0947	0.0833	0.0785
paired	pwc	easy_to_understand	gemini-3.5-flash-lite	500	3	2	0.1465	0.1425	0.1506	0.1376	0.1335
paired	pwc	easy_to_understand	claude-haiku-4-5	500	67	34	0.1337	0.1244	0.1434	0.1221	0.1138
paired	pwc	easy_to_understand	gpt-5.6-luna	500	62	15	0.0893	0.0840	0.0947	0.0641	0.0596
paired	pwc	authoritative	gemini-3.5-flash-lite	500	3	1	0.1465	0.1425	0.1506	0.1427	0.1388
paired	pwc	authoritative	claude-haiku-4-5	500	67	25	0.1337	0.1244	0.1434	0.1199	0.1114
paired	pwc	authoritative	gpt-5.6-luna	500	62	5	0.0893	0.0840	0.0947	0.0484	0.0441
paired	pwc	unique_words	gemini-3.5-flash-lite	500	3	1	0.1465	0.1425	0.1506	0.1416	0.1377
paired	pwc	unique_words	claude-haiku-4-5	500	67	9	0.1337	0.1244	0.1434	0.0700	0.0627
paired	pwc	unique_words	gpt-5.6-luna	500	62	2	0.0893	0.0840	0.0947	0.0439	0.0399
paired	pwc	technical_terms	gemini-3.5-flash-lite	500	3	2	0.1465	0.1425	0.1506	0.1514	0.1477
paired	pwc	technical_terms	claude-haiku-4-5	500	67	29	0.1337	0.1244	0.1434	0.1329	0.1239
paired	pwc	technical_terms	gpt-5.6-luna	500	62	15	0.0893	0.0840	0.0947	0.0764	0.0715
paired	pwc	keyword_stuffing	gemini-3.5-flash-lite	500	3	1	0.1465	0.1425	0.1506	0.1307	0.1263
paired	pwc	keyword_stuffing	claude-haiku-4-5	500	67	21	0.1337	0.1244	0.1434	0.0980	0.0895
paired	pwc	keyword_stuffing	gpt-5.6-luna	500	62	9	0.0893	0.0840	0.0947	0.0528	0.0484
baseline_pos	pwc	cite_sources	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1574	0.1536
baseline_pos	pwc	cite_sources	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1629	0.1525
baseline_pos	pwc	cite_sources	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0923	0.0870
baseline_pos	pwc	quotation_addition	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1557	0.1523
baseline_pos	pwc	quotation_addition	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1619	0.1517
baseline_pos	pwc	quotation_addition	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0967	0.0914
baseline_pos	pwc	statistics_addition	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1558	0.1517
baseline_pos	pwc	statistics_addition	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1640	0.1530
baseline_pos	pwc	statistics_addition	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0412	0.0360
baseline_pos	pwc	fluency_optimization	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1576	0.1539
baseline_pos	pwc	fluency_optimization	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1774	0.1666
baseline_pos	pwc	fluency_optimization	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0991	0.0941
baseline_pos	pwc	easy_to_understand	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1462	0.1420
baseline_pos	pwc	easy_to_understand	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1425	0.1334
baseline_pos	pwc	easy_to_understand	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0758	0.0709
baseline_pos	pwc	authoritative	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1504	0.1463
baseline_pos	pwc	authoritative	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1414	0.1321
baseline_pos	pwc	authoritative	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0583	0.0534
baseline_pos	pwc	unique_words	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1494	0.1458
baseline_pos	pwc	unique_words	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.0843	0.0760
baseline_pos	pwc	unique_words	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0535	0.0488
baseline_pos	pwc	technical_terms	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1592	0.1554
baseline_pos	pwc	technical_terms	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1583	0.1482
baseline_pos	pwc	technical_terms	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0906	0.0852
baseline_pos	pwc	keyword_stuffing	gemini-3.5-flash-lite	402	0	0	0.1560	0.1521	0.1598	0.1408	0.1363
baseline_pos	pwc	keyword_stuffing	claude-haiku-4-5	402	0	0	0.1621	0.1525	0.1720	0.1159	0.1067
baseline_pos	pwc	keyword_stuffing	gpt-5.6-luna	402	0	0	0.1072	0.1019	0.1124	0.0630	0.0580
paired	wc	cite_sources	gemini-3.5-flash-lite	500	3	2	0.1986	0.1936	0.2037	0.2048	0.2002
paired	wc	cite_sources	claude-haiku-4-5	500	67	32	0.1844	0.1717	0.1972	0.1890	0.1760
paired	wc	cite_sources	gpt-5.6-luna	500	62	9	0.1356	0.1279	0.1435	0.1177	0.1104
paired	wc	quotation_addition	gemini-3.5-flash-lite	500	3	2	0.1986	0.1936	0.2037	0.2011	0.1967
paired	wc	quotation_addition	claude-haiku-4-5	500	67	31	0.1844	0.1717	0.1972	0.1868	0.1737
paired	wc	quotation_addition	gpt-5.6-luna	500	62	9	0.1356	0.1279	0.1435	0.1232	0.1158
paired	wc	statistics_addition	gemini-3.5-flash-lite	500	3	1	0.1986	0.1936	0.2037	0.2007	0.1954
paired	wc	statistics_addition	claude-haiku-4-5	500	67	33	0.1844	0.1717	0.1972	0.1997	0.1858
paired	wc	statistics_addition	gpt-5.6-luna	500	62	3	0.1356	0.1279	0.1435	0.0517	0.0452
paired	wc	fluency_optimization	gemini-3.5-flash-lite	500	3	1	0.1986	0.1936	0.2037	0.2044	0.1998
paired	wc	fluency_optimization	claude-haiku-4-5	500	67	46	0.1844	0.1717	0.1972	0.2134	0.2000
paired	wc	fluency_optimization	gpt-5.6-luna	500	62	13	0.1356	0.1279	0.1435	0.1269	0.1197
paired	wc	easy_to_understand	gemini-3.5-flash-lite	500	3	2	0.1986	0.1936	0.2037	0.1887	0.1836
paired	wc	easy_to_understand	claude-haiku-4-5	500	67	34	0.1844	0.1717	0.1972	0.1713	0.1599
paired	wc	easy_to_understand	gpt-5.6-luna	500	62	15	0.1356	0.1279	0.1435	0.0986	0.0920
paired	wc	authoritative	gemini-3.5-flash-lite	500	3	1	0.1986	0.1936	0.2037	0.1936	0.1888
paired	wc	authoritative	claude-haiku-4-5	500	67	25	0.1844	0.1717	0.1972	0.1645	0.1533
paired	wc	authoritative	gpt-5.6-luna	500	62	5	0.1356	0.1279	0.1435	0.0745	0.0680
paired	wc	unique_words	gemini-3.5-flash-lite	500	3	1	0.1986	0.1936	0.2037	0.1917	0.1869
paired	wc	unique_words	claude-haiku-4-5	500	67	9	0.1844	0.1717	0.1972	0.0947	0.0854
paired	wc	unique_words	gpt-5.6-luna	500	62	2	0.1356	0.1279	0.1435	0.0678	0.0618
paired	wc	technical_terms	gemini-3.5-flash-lite	500	3	2	0.1986	0.1936	0.2037	0.2062	0.2016
paired	wc	technical_terms	claude-haiku-4-5	500	67	29	0.1844	0.1717	0.1972	0.1823	0.1700
paired	wc	technical_terms	gpt-5.6-luna	500	62	15	0.1356	0.1279	0.1435	0.1169	0.1097
paired	wc	keyword_stuffing	gemini-3.5-flash-lite	500	3	1	0.1986	0.1936	0.2037	0.1741	0.1687
paired	wc	keyword_stuffing	claude-haiku-4-5	500	67	21	0.1844	0.1717	0.1972	0.1300	0.1189
paired	wc	keyword_stuffing	gpt-5.6-luna	500	62	9	0.1356	0.1279	0.1435	0.0795	0.0730
baseline_pos	wc	cite_sources	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2149	0.2103
baseline_pos	wc	cite_sources	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2235	0.2094
baseline_pos	wc	cite_sources	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.1407	0.1329
baseline_pos	wc	quotation_addition	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2129	0.2089

continued on next page

Table 18 – continued from previous page

set	metric	technique	engine	n	n_baseline_zero	n_indefinido	base.	baseline_ci_lo	baseline_ci_hi	tech.	tecnica_ci_lo
baseline_pos	wc	quotation_addition	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2228	0.2088
baseline_pos	wc	quotation_addition	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.1472	0.1396
baseline_pos	wc	statistics_addition	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2135	0.2084
baseline_pos	wc	statistics_addition	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2353	0.2197
baseline_pos	wc	statistics_addition	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.0628	0.0551
baseline_pos	wc	fluency_optimization	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2160	0.2116
baseline_pos	wc	fluency_optimization	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2485	0.2337
baseline_pos	wc	fluency_optimization	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.1507	0.1436
baseline_pos	wc	easy_to_understand	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.1996	0.1945
baseline_pos	wc	easy_to_understand	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2000	0.1875
baseline_pos	wc	easy_to_understand	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.1161	0.1089
baseline_pos	wc	authoritative	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2043	0.1995
baseline_pos	wc	authoritative	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.1941	0.1817
baseline_pos	wc	authoritative	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.0895	0.0821
baseline_pos	wc	unique_words	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2029	0.1987
baseline_pos	wc	unique_words	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.1143	0.1035
baseline_pos	wc	unique_words	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.0823	0.0755
baseline_pos	wc	technical_terms	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.2172	0.2125
baseline_pos	wc	technical_terms	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.2171	0.2039
baseline_pos	wc	technical_terms	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.1383	0.1305
baseline_pos	wc	keyword_stuffing	gemini-3.5-flash-lite	402	0	0	0.2116	0.2070	0.2162	0.1870	0.1816
baseline_pos	wc	keyword_stuffing	claude-haiku-4-5	402	0	0	0.2231	0.2101	0.2367	0.1542	0.1420
baseline_pos	wc	keyword_stuffing	gpt-5.6-luna	402	0	0	0.1627	0.1552	0.1701	0.0945	0.0874

Table 19: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Part 2 of 5** of the complete table; the identifying columns repeat in every part.

set	metric	technique	engine	n	tecnica_ci_hi	mean %	melhoria_media_ci_lo	melhoria_media_ci_hi	median %
paired	pwc	cite_sources	gemini-3.5-flash-lite	500	0.1543	15.32	7.157	26.78	1.740
paired	pwc	cite_sources	claude-haiku-4-5	500	0.1477	32.87	13.55	58.49	0
paired	pwc	cite_sources	gpt-5.6-luna	500	0.0824	-3.717	-12.35	8.352	-7.259
paired	pwc	quotation_addition	gemini-3.5-flash-lite	500	0.1510	11.75	3.428	26.04	0.5363
paired	pwc	quotation_addition	claude-haiku-4-5	500	0.1453	47.36	8.343	115.2	0
paired	pwc	quotation_addition	gpt-5.6-luna	500	0.0859	-4.181	-9.523	1.667	-3.974
paired	pwc	statistics_addition	gemini-3.5-flash-lite	500	0.1506	10.98	4.761	18.28	-1.324
paired	pwc	statistics_addition	claude-haiku-4-5	500	0.1495	45.76	20.60	78.00	0
paired	pwc	statistics_addition	gpt-5.6-luna	500	0.0384	-57.69	-62.40	-52.77	-74.56
paired	pwc	fluency_optimization	gemini-3.5-flash-lite	500	0.1531	10.71	5.459	17.26	0.6499
paired	pwc	fluency_optimization	claude-haiku-4-5	500	0.1625	87.10	28.04	194.4	5.041
paired	pwc	fluency_optimization	gpt-5.6-luna	500	0.0883	7.757	-1.050	20.07	-3.927
paired	pwc	easy_to_understand	gemini-3.5-flash-lite	500	0.1416	5.686	-3.885	20.33	-5.255
paired	pwc	easy_to_understand	claude-haiku-4-5	500	0.1305	26.36	9.457	47.44	-7.220
paired	pwc	easy_to_understand	gpt-5.6-luna	500	0.0685	-17.77	-23.93	-11.03	-27.35
paired	pwc	authoritative	gemini-3.5-flash-lite	500	0.1468	8.934	0.5457	22.03	-2.343
paired	pwc	authoritative	claude-haiku-4-5	500	0.1286	41.25	4.955	105.4	-3.846
paired	pwc	authoritative	gpt-5.6-luna	500	0.0528	-40.69	-45.61	-35.40	-43.29
paired	pwc	unique_words	gemini-3.5-flash-lite	500	0.1454	1.810	-1.731	5.695	-2.062
paired	pwc	unique_words	claude-haiku-4-5	500	0.0775	-33.09	-48.39	-7.586	-46.46
paired	pwc	unique_words	gpt-5.6-luna	500	0.0480	-47.80	-51.39	-44.20	-50.02
paired	pwc	technical_terms	gemini-3.5-flash-lite	500	0.1553	11.70	7.351	16.68	2.526
paired	pwc	technical_terms	claude-haiku-4-5	500	0.1425	29.31	6.850	66.11	-0.7657
paired	pwc	technical_terms	gpt-5.6-luna	500	0.0814	-1.972	-10.50	8.526	-8.906
paired	pwc	keyword_stuffing	gemini-3.5-flash-lite	500	0.1350	-4.003	-9.010	1.992	-9.751
paired	pwc	keyword_stuffing	claude-haiku-4-5	500	0.1067	-8.163	-21.22	8.817	-25.24
paired	pwc	keyword_stuffing	gpt-5.6-luna	500	0.0574	-37.20	-41.93	-32.19	-40.71
baseline_pos	pwc	cite_sources	gemini-3.5-flash-lite	402	0.1612	4.308	1.616	7.254	0.4816
baseline_pos	pwc	cite_sources	claude-haiku-4-5	402	0.1739	31.22	11.74	58.31	-2.329
baseline_pos	pwc	cite_sources	gpt-5.6-luna	402	0.0976	-2.774	-12.88	11.89	-12.08
baseline_pos	pwc	quotation_addition	gemini-3.5-flash-lite	402	0.1591	3.257	0.7748	6.072	0.0055
baseline_pos	pwc	quotation_addition	claude-haiku-4-5	402	0.1724	53.68	8.591	130.6	-2.051
baseline_pos	pwc	quotation_addition	gpt-5.6-luna	402	0.1020	-5.630	-11.44	0.6375	-9.551
baseline_pos	pwc	statistics_addition	gemini-3.5-flash-lite	402	0.1600	3.607	0.5158	6.996	-1.449
baseline_pos	pwc	statistics_addition	claude-haiku-4-5	402	0.1753	51.18	22.84	88.39	0.5858
baseline_pos	pwc	statistics_addition	gpt-5.6-luna	402	0.0468	-63.65	-68.85	-57.88	-81.53
baseline_pos	pwc	fluency_optimization	gemini-3.5-flash-lite	402	0.1612	4.641	1.928	7.834	0.1682
baseline_pos	pwc	fluency_optimization	claude-haiku-4-5	402	0.1888	90.77	25.46	209.8	7.886
baseline_pos	pwc	fluency_optimization	gpt-5.6-luna	402	0.1040	8.163	-1.994	22.24	-6.891
baseline_pos	pwc	easy_to_understand	gemini-3.5-flash-lite	402	0.1504	-3.134	-6.491	0.9800	-5.447
baseline_pos	pwc	easy_to_understand	claude-haiku-4-5	402	0.1514	17.89	3.969	34.34	-10.66
baseline_pos	pwc	easy_to_understand	gpt-5.6-luna	402	0.0807	-21.57	-27.95	-14.03	-31.37
baseline_pos	pwc	authoritative	gemini-3.5-flash-lite	402	0.1543	0.4930	-2.887	4.467	-3.132
baseline_pos	pwc	authoritative	claude-haiku-4-5	402	0.1507	44.07	2.147	118.0	-8.092
baseline_pos	pwc	authoritative	gpt-5.6-luna	402	0.0634	-44.42	-49.94	-38.18	-48.68
baseline_pos	pwc	unique_words	gemini-3.5-flash-lite	402	0.1531	-1.196	-3.746	1.909	-3.218
baseline_pos	pwc	unique_words	claude-haiku-4-5	402	0.0930	-36.04	-53.95	-5.129	-54.12
baseline_pos	pwc	unique_words	gpt-5.6-luna	402	0.0582	-53.49	-57.10	-49.67	-55.37
baseline_pos	pwc	technical_terms	gemini-3.5-flash-lite	402	0.1632	5.676	2.902	9.028	1.920
baseline_pos	pwc	technical_terms	claude-haiku-4-5	402	0.1686	33.03	7.480	73.71	-4.371
baseline_pos	pwc	technical_terms	gpt-5.6-luna	402	0.0959	-6.815	-14.15	1.543	-12.41
baseline_pos	pwc	keyword_stuffing	gemini-3.5-flash-lite	402	0.1452	-7.469	-10.68	-3.954	-9.103
baseline_pos	pwc	keyword_stuffing	claude-haiku-4-5	402	0.1257	-5.305	-20.68	13.82	-28.62
baseline_pos	pwc	keyword_stuffing	gpt-5.6-luna	402	0.0682	-40.96	-46.19	-35.13	-45.31
paired	wc	cite_sources	gemini-3.5-flash-lite	500	0.2096	10.73	5.982	16.68	2.175

continued on next page

Table 19 – continued from previous page

set	metric	technique	engine	n	tecnica_ci_hi	mean %	melhoria_media_ci_lo	melhoria_media_ci_hi	median %
paired	wc	cite_sources	claude-haiku-4-5	500	0.2023	26.28	9.379	50.23	0
paired	wc	cite_sources	gpt-5.6-luna	500	0.1254	-5.012	-12.33	4.486	-7.461
paired	wc	quotation_addition	gemini-3.5-flash-lite	500	0.2056	7.747	3.145	14.40	0.2370
paired	wc	quotation_addition	claude-haiku-4-5	500	0.2001	49.05	4.247	131.1	0
paired	wc	quotation_addition	gpt-5.6-luna	500	0.1308	-4.149	-9.343	1.652	-3.520
paired	wc	statistics_addition	gemini-3.5-flash-lite	500	0.2061	8.734	4.012	14.11	0
paired	wc	statistics_addition	claude-haiku-4-5	500	0.2140	39.55	19.32	64.21	0
paired	wc	statistics_addition	gpt-5.6-luna	500	0.0585	-57.37	-62.08	-52.31	-72.28
paired	wc	fluency_optimization	gemini-3.5-flash-lite	500	0.2090	9.680	5.432	14.70	0.5522
paired	wc	fluency_optimization	claude-haiku-4-5	500	0.2275	79.94	25.80	178.1	6.660
paired	wc	fluency_optimization	gpt-5.6-luna	500	0.1342	10.10	-1.449	28.60	-2.846
paired	wc	easy_to_understand	gemini-3.5-flash-lite	500	0.1937	1.503	-3.818	8.502	-3.068
paired	wc	easy_to_understand	claude-haiku-4-5	500	0.1829	18.16	5.132	34.68	-6.654
paired	wc	easy_to_understand	gpt-5.6-luna	500	0.1053	-17.54	-24.02	-9.689	-26.30
paired	wc	authoritative	gemini-3.5-flash-lite	500	0.1985	4.642	-0.4681	11.36	-2.120
paired	wc	authoritative	claude-haiku-4-5	500	0.1759	44.42	1.052	123.5	-5.682
paired	wc	authoritative	gpt-5.6-luna	500	0.0811	-39.38	-44.99	-32.70	-41.28
paired	wc	unique_words	gemini-3.5-flash-lite	500	0.1964	0.3760	-2.607	3.681	-1.665
paired	wc	unique_words	claude-haiku-4-5	500	0.1044	-35.89	-49.65	-12.69	-45.94
paired	wc	unique_words	gpt-5.6-luna	500	0.0739	-47.23	-50.77	-43.70	-48.41
paired	wc	technical_terms	gemini-3.5-flash-lite	500	0.2111	9.851	6.344	13.81	2.534
paired	wc	technical_terms	claude-haiku-4-5	500	0.1953	35.12	3.791	90.95	-2.291
paired	wc	technical_terms	gpt-5.6-luna	500	0.1243	-1.983	-10.52	8.428	-8.004
paired	wc	keyword_stuffing	gemini-3.5-flash-lite	500	0.1794	-7.693	-11.37	-3.590	-8.423
paired	wc	keyword_stuffing	claude-haiku-4-5	500	0.1412	-10.22	-25.77	13.95	-26.81
paired	wc	keyword_stuffing	gpt-5.6-luna	500	0.0862	-37.67	-42.06	-33.18	-39.22
baseline_pos	wc	cite_sources	gemini-3.5-flash-lite	402	0.2197	4.216	1.856	6.955	0.6922
baseline_pos	wc	cite_sources	claude-haiku-4-5	402	0.2383	26.24	8.460	52.84	-0.8432
baseline_pos	wc	cite_sources	gpt-5.6-luna	402	0.1485	-4.097	-12.67	7.179	-11.33
baseline_pos	wc	quotation_addition	gemini-3.5-flash-lite	402	0.2169	3.342	1.159	5.841	0
baseline_pos	wc	quotation_addition	claude-haiku-4-5	402	0.2368	56.38	4.745	149.8	-0.5784
baseline_pos	wc	quotation_addition	gpt-5.6-luna	402	0.1548	-5.599	-11.33	0.7606	-8.547
baseline_pos	wc	statistics_addition	gemini-3.5-flash-lite	402	0.2191	3.786	0.9308	6.980	-0.1738
baseline_pos	wc	statistics_addition	claude-haiku-4-5	402	0.2512	44.52	22.11	72.77	0.7699
baseline_pos	wc	statistics_addition	gpt-5.6-luna	402	0.0712	-63.23	-68.41	-57.39	-80.09
baseline_pos	wc	fluency_optimization	gemini-3.5-flash-lite	402	0.2205	4.799	2.433	7.624	0.3651
baseline_pos	wc	fluency_optimization	claude-haiku-4-5	402	0.2640	84.57	24.52	193.3	9.161
baseline_pos	wc	fluency_optimization	gpt-5.6-luna	402	0.1579	11.04	-2.510	32.41	-6.529
baseline_pos	wc	easy_to_understand	gemini-3.5-flash-lite	402	0.2046	-3.512	-6.349	-0.2336	-3.247
baseline_pos	wc	easy_to_understand	claude-haiku-4-5	402	0.2126	12.34	1.173	25.19	-10.97
baseline_pos	wc	easy_to_understand	gpt-5.6-luna	402	0.1233	-21.23	-28.10	-12.25	-30.66
baseline_pos	wc	authoritative	gemini-3.5-flash-lite	402	0.2090	-0.3811	-3.256	2.868	-2.805
baseline_pos	wc	authoritative	claude-haiku-4-5	402	0.2066	48.88	-1.203	139.9	-9.352
baseline_pos	wc	authoritative	gpt-5.6-luna	402	0.0969	-42.97	-49.33	-35.14	-47.23
baseline_pos	wc	unique_words	gemini-3.5-flash-lite	402	0.2072	-1.619	-3.888	1.110	-2.051
baseline_pos	wc	unique_words	claude-haiku-4-5	402	0.1257	-39.59	-55.41	-12.14	-52.79
baseline_pos	wc	unique_words	gpt-5.6-luna	402	0.0892	-52.78	-56.33	-49.12	-53.98
baseline_pos	wc	technical_terms	gemini-3.5-flash-lite	402	0.2223	5.428	2.948	8.521	1.814
baseline_pos	wc	technical_terms	claude-haiku-4-5	402	0.2313	41.22	5.015	104.6	-5.965
baseline_pos	wc	technical_terms	gpt-5.6-luna	402	0.1460	-6.098	-13.80	3.151	-12.28
baseline_pos	wc	keyword_stuffing	gemini-3.5-flash-lite	402	0.1923	-9.789	-12.74	-6.605	-8.575
baseline_pos	wc	keyword_stuffing	claude-haiku-4-5	402	0.1671	-7.336	-25.68	20.61	-30.58
baseline_pos	wc	keyword_stuffing	gpt-5.6-luna	402	0.1020	-41.56	-46.35	-36.41	-43.84

Table 20: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Part 3 of 5** of the complete table; the identifying columns repeat in every part.

set	metric	technique	engine	n	CI lo	CI hi	CI excl. 0 (mean)	CI excl. 0	p
paired	pwc	cite_sources	gemini-3.5-flash-lite	500	0.2001	3.796	yes (+)	yes (+)	0.0256
paired	pwc	cite_sources	claude-haiku-4-5	500	-2.604	0	yes (+)	no	1.000
paired	pwc	cite_sources	gpt-5.6-luna	500	-12.92	-3.509	no	yes (−)	0.0002
paired	pwc	quotation_addition	gemini-3.5-flash-lite	500	-0.6521	2.132	yes (+)	no	0.3550
paired	pwc	quotation_addition	claude-haiku-4-5	500	-2.361	0	yes (+)	no	1.000
paired	pwc	quotation_addition	gpt-5.6-luna	500	-8.520	-0.5781	no	yes (−)	0.0304
paired	pwc	statistics_addition	gemini-3.5-flash-lite	500	-3.114	0.5866	yes (+)	no	0.2106
paired	pwc	statistics_addition	claude-haiku-4-5	500	-3.469	0.0803	yes (+)	no	1.000
paired	pwc	statistics_addition	gpt-5.6-luna	500	-82.39	-65.13	yes (−)	yes (−)	0.0001
paired	pwc	fluency_optimization	gemini-3.5-flash-lite	500	-0.3601	2.313	yes (+)	no	0.2940
paired	pwc	fluency_optimization	claude-haiku-4-5	500	0.8291	11.35	yes (+)	yes (+)	0.0276
paired	pwc	fluency_optimization	gpt-5.6-luna	500	-6.257	0	no	no	0.0880
paired	pwc	easy_to_understand	gemini-3.5-flash-lite	500	-7.573	-3.459	no	yes (−)	0.0001
paired	pwc	easy_to_understand	claude-haiku-4-5	500	-11.30	-1.385	yes (+)	yes (−)	0.0090
paired	pwc	easy_to_understand	gpt-5.6-luna	500	-31.36	-20.95	yes (−)	yes (−)	0.0001
paired	pwc	authoritative	gemini-3.5-flash-lite	500	-4.193	0	yes (+)	yes (−)	0.0404
paired	pwc	authoritative	claude-haiku-4-5	500	-8.229	0	yes (+)	no	0.1024
paired	pwc	authoritative	gpt-5.6-luna	500	-49.07	-36.16	yes (−)	yes (−)	0.0001
paired	pwc	unique_words	gemini-3.5-flash-lite	500	-4.230	-0.2078	no	yes (−)	0.0330
paired	pwc	unique_words	claude-haiku-4-5	500	-54.38	-41.30	yes (−)	yes (−)	0.0001
paired	pwc	unique_words	gpt-5.6-luna	500	-54.82	-44.07	yes (−)	yes (−)	0.0001
paired	pwc	technical_terms	gemini-3.5-flash-lite	500	0.9487	4.084	yes (+)	yes (+)	0.0020
paired	pwc	technical_terms	claude-haiku-4-5	500	-5.769	0	yes (+)	no	0.6636
paired	pwc	technical_terms	gpt-5.6-luna	500	-11.79	-4.669	no	yes (−)	0.0002
paired	pwc	keyword_stuffing	gemini-3.5-flash-lite	500	-11.66	-5.590	no	yes (−)	0.0001

continued on next page

Table 20 – continued from previous page

set	metric	technique	engine	n	CI lo	CI hi	CI excl. 0 (mean)	CI excl. 0	p
paired	pwc	keyword_stuffing	claude-haiku-4-5	500	-30.06	-19.62	no	yes (−)	0.0001
paired	pwc	keyword_stuffing	gpt-5.6-luna	500	-44.44	-33.19	yes (−)	yes (−)	0.0001
baseline_pos	pwc	cite_sources	gemini-3.5-flash-lite	402	-1.216	2.409	yes (+)	no	0.4628
baseline_pos	pwc	cite_sources	claude-haiku-4-5	402	-5.646	2.169	yes (+)	no	0.4028
baseline_pos	pwc	cite_sources	gpt-5.6-luna	402	-17.13	-7.526	no	yes (−)	0.0001
baseline_pos	pwc	quotation_addition	gemini-3.5-flash-lite	402	-1.485	1.591	yes (+)	no	0.9490
baseline_pos	pwc	quotation_addition	claude-haiku-4-5	402	-5.710	1.594	yes (+)	no	0.1836
baseline_pos	pwc	quotation_addition	gpt-5.6-luna	402	-12.81	-5.194	no	yes (−)	0.0001
baseline_pos	pwc	statistics_addition	gemini-3.5-flash-lite	402	-3.560	0.5726	yes (+)	no	0.1548
baseline_pos	pwc	statistics_addition	claude-haiku-4-5	402	-8.572	4.280	yes (+)	no	0.8812
baseline_pos	pwc	statistics_addition	gpt-5.6-luna	402	-91.78	-73.88	yes (−)	yes (−)	0.0001
baseline_pos	pwc	fluency_optimization	gemini-3.5-flash-lite	402	-1.377	1.632	yes (+)	no	0.7278
baseline_pos	pwc	fluency_optimization	claude-haiku-4-5	402	2.014	13.18	yes (+)	yes (+)	0.0026
baseline_pos	pwc	fluency_optimization	gpt-5.6-luna	402	-9.592	-4.514	no	yes (−)	0.0001
baseline_pos	pwc	easy_to_understand	gemini-3.5-flash-lite	402	-7.600	-3.683	no	yes (−)	0.0001
baseline_pos	pwc	easy_to_understand	claude-haiku-4-5	402	-15.49	-7.171	yes (+)	yes (−)	0.0001
baseline_pos	pwc	easy_to_understand	gpt-5.6-luna	402	-34.72	-27.50	yes (−)	yes (−)	0.0001
baseline_pos	pwc	authoritative	gemini-3.5-flash-lite	402	-4.882	-0.9540	no	yes (−)	0.0026
baseline_pos	pwc	authoritative	claude-haiku-4-5	402	-15.98	-3.878	yes (+)	yes (−)	0.0002
baseline_pos	pwc	authoritative	gpt-5.6-luna	402	-55.75	-43.39	yes (−)	yes (−)	0.0001
baseline_pos	pwc	unique_words	gemini-3.5-flash-lite	402	-5.183	-1.001	no	yes (−)	0.0026
baseline_pos	pwc	unique_words	claude-haiku-4-5	402	-65.80	-46.73	yes (−)	yes (−)	0.0001
baseline_pos	pwc	unique_words	gpt-5.6-luna	402	-60.05	-51.26	yes (−)	yes (−)	0.0001
baseline_pos	pwc	technical_terms	gemini-3.5-flash-lite	402	0.0067	3.814	yes (+)	yes (+)	0.0386
baseline_pos	pwc	technical_terms	claude-haiku-4-5	402	-8.163	-0.0948	yes (+)	yes (−)	0.0434
baseline_pos	pwc	technical_terms	gpt-5.6-luna	402	-16.98	-9.483	no	yes (−)	0.0001
baseline_pos	pwc	keyword_stuffing	gemini-3.5-flash-lite	402	-11.53	-5.459	yes (−)	yes (−)	0.0001
baseline_pos	pwc	keyword_stuffing	claude-haiku-4-5	402	-34.36	-23.13	no	yes (−)	0.0001
baseline_pos	pwc	keyword_stuffing	gpt-5.6-luna	402	-50.11	-41.07	yes (−)	yes (−)	0.0001
paired	wc	cite_sources	gemini-3.5-flash-lite	500	0.3590	3.657	yes (+)	yes (+)	0.0032
paired	wc	cite_sources	claude-haiku-4-5	500	-1.407	0	yes (+)	no	1.000
paired	wc	cite_sources	gpt-5.6-luna	500	-12.19	-3.190	no	yes (−)	0.0001
paired	wc	quotation_addition	gemini-3.5-flash-lite	500	-0.0635	1.558	yes (+)	no	0.5146
paired	wc	quotation_addition	claude-haiku-4-5	500	-1.590	0	yes (+)	no	1.000
paired	wc	quotation_addition	gpt-5.6-luna	500	-7.852	-0.2283	no	yes (−)	0.0394
paired	wc	statistics_addition	gemini-3.5-flash-lite	500	-2.104	1.275	yes (+)	no	1.000
paired	wc	statistics_addition	claude-haiku-4-5	500	-2.226	0.6149	yes (+)	no	1.000
paired	wc	statistics_addition	gpt-5.6-luna	500	-81.85	-63.36	yes (−)	yes (−)	0.0001
paired	wc	fluency_optimization	gemini-3.5-flash-lite	500	0	2.201	yes (+)	no	0.1850
paired	wc	fluency_optimization	claude-haiku-4-5	500	2.258	11.74	yes (+)	yes (+)	0.0022
paired	wc	fluency_optimization	gpt-5.6-luna	500	-6.384	0	no	no	0.1070
paired	wc	easy_to_understand	gemini-3.5-flash-lite	500	-4.553	-1.383	no	yes (−)	0.0001
paired	wc	easy_to_understand	claude-haiku-4-5	500	-11.16	-1.709	yes (+)	yes (−)	0.0074
paired	wc	easy_to_understand	gpt-5.6-luna	500	-29.96	-20.53	yes (−)	yes (−)	0.0001
paired	wc	authoritative	gemini-3.5-flash-lite	500	-3.733	0	no	no	0.0518
paired	wc	authoritative	claude-haiku-4-5	500	-9.675	-0.5388	yes (+)	yes (−)	0.0244
paired	wc	authoritative	gpt-5.6-luna	500	-47.09	-35.86	yes (−)	yes (−)	0.0001
paired	wc	unique_words	gemini-3.5-flash-lite	500	-3.249	-0.5139	no	yes (−)	0.0026
paired	wc	unique_words	claude-haiku-4-5	500	-52.12	-40.12	yes (−)	yes (−)	0.0001
paired	wc	unique_words	gpt-5.6-luna	500	-52.73	-41.88	yes (−)	yes (−)	0.0001
paired	wc	technical_terms	gemini-3.5-flash-lite	500	1.072	3.969	yes (+)	yes (+)	0.0016
paired	wc	technical_terms	claude-haiku-4-5	500	-6.411	0	yes (+)	no	0.3338
paired	wc	technical_terms	gpt-5.6-luna	500	-11.67	-4.073	no	yes (−)	0.0006
paired	wc	keyword_stuffing	gemini-3.5-flash-lite	500	-10.93	-6.409	yes (−)	yes (−)	0.0001
paired	wc	keyword_stuffing	claude-haiku-4-5	500	-31.94	-21.27	no	yes (−)	0.0001
paired	wc	keyword_stuffing	gpt-5.6-luna	500	-43.78	-34.81	yes (−)	yes (−)	0.0001
baseline_pos	wc	cite_sources	gemini-3.5-flash-lite	402	0	2.394	yes (+)	no	0.1976
baseline_pos	wc	cite_sources	claude-haiku-4-5	402	-3.608	1.705	yes (+)	no	0.7444
baseline_pos	wc	cite_sources	gpt-5.6-luna	402	-15.91	-7.461	no	yes (−)	0.0001
baseline_pos	wc	quotation_addition	gemini-3.5-flash-lite	402	-1.009	1.336	yes (+)	no	1.000
baseline_pos	wc	quotation_addition	claude-haiku-4-5	402	-5.573	2.594	yes (+)	no	0.6602
baseline_pos	wc	quotation_addition	gpt-5.6-luna	402	-13.10	-4.203	no	yes (−)	0.0001
baseline_pos	wc	statistics_addition	gemini-3.5-flash-lite	402	-2.404	1.065	yes (+)	no	0.5192
baseline_pos	wc	statistics_addition	claude-haiku-4-5	402	-5.900	6.504	yes (+)	no	0.5268
baseline_pos	wc	statistics_addition	gpt-5.6-luna	402	-89.60	-71.94	yes (−)	yes (−)	0.0001
baseline_pos	wc	fluency_optimization	gemini-3.5-flash-lite	402	-0.7878	1.599	yes (+)	no	0.5148
baseline_pos	wc	fluency_optimization	claude-haiku-4-5	402	4.092	13.22	yes (+)	yes (+)	0.0001
baseline_pos	wc	fluency_optimization	gpt-5.6-luna	402	-9.166	-3.566	no	yes (−)	0.0002
baseline_pos	wc	easy_to_understand	gemini-3.5-flash-lite	402	-5.119	-2.142	yes (−)	yes (−)	0.0001
baseline_pos	wc	easy_to_understand	claude-haiku-4-5	402	-14.70	-6.534	yes (+)	yes (−)	0.0001
baseline_pos	wc	easy_to_understand	gpt-5.6-luna	402	-35.17	-26.83	yes (−)	yes (−)	0.0001
baseline_pos	wc	authoritative	gemini-3.5-flash-lite	402	-4.209	-1.138	no	yes (−)	0.0060
baseline_pos	wc	authoritative	claude-haiku-4-5	402	-14.44	-5.477	no	yes (−)	0.0001
baseline_pos	wc	authoritative	gpt-5.6-luna	402	-55.01	-42.03	yes (−)	yes (−)	0.0001
baseline_pos	wc	unique_words	gemini-3.5-flash-lite	402	-3.869	-0.7219	no	yes (−)	0.0008
baseline_pos	wc	unique_words	claude-haiku-4-5	402	-63.23	-46.82	yes (−)	yes (−)	0.0001
baseline_pos	wc	unique_words	gpt-5.6-luna	402	-57.74	-49.70	yes (−)	yes (−)	0.0001
baseline_pos	wc	technical_terms	gemini-3.5-flash-lite	402	0.0578	3.449	yes (+)	yes (+)	0.0192
baseline_pos	wc	technical_terms	claude-haiku-4-5	402	-9.002	-1.378	yes (+)	yes (−)	0.0102
baseline_pos	wc	technical_terms	gpt-5.6-luna	402	-16.32	-8.873	no	yes (−)	0.0001
baseline_pos	wc	keyword_stuffing	gemini-3.5-flash-lite	402	-10.89	-6.703	yes (−)	yes (−)	0.0001
baseline_pos	wc	keyword_stuffing	claude-haiku-4-5	402	-36.97	-25.70	no	yes (−)	0.0001
baseline_pos	wc	keyword_stuffing	gpt-5.6-luna	402	-50.30	-39.93	yes (−)	yes (−)	0.0001

Table 21: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Part 4 of 5** of the complete table; the identifying columns repeat in every part.

set	metric	technique	engine	n	p (Holm)	sig. (Holm)	p (BH)	sig. (BH)	n_citacao_zerada
paired	pwc	cite_sources	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	cite_sources	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	29
paired	pwc	cite_sources	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	30
paired	pwc	quotation_addition	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	quotation_addition	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	25
paired	pwc	quotation_addition	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	26
paired	pwc	statistics_addition	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	0
paired	pwc	statistics_addition	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	37
paired	pwc	statistics_addition	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	195
paired	pwc	fluency_optimization	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	fluency_optimization	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	19
paired	pwc	fluency_optimization	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	10
paired	pwc	easy_to_understand	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	easy_to_understand	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	30
paired	pwc	easy_to_understand	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	37
paired	pwc	authoritative	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	2
paired	pwc	authoritative	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	33
paired	pwc	authoritative	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	95
paired	pwc	unique_words	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	3
paired	pwc	unique_words	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	125
paired	pwc	unique_words	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	101
paired	pwc	technical_terms	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	technical_terms	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	33
paired	pwc	technical_terms	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	31
paired	pwc	keyword_stuffing	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	pwc	keyword_stuffing	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	78
paired	pwc	keyword_stuffing	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	75
baseline_pos	pwc	cite_sources	gemini-3.5-flash-lite	402	1.000	no	0.5206	não (BH)	0
baseline_pos	pwc	cite_sources	claude-haiku-4-5	402	1.000	no	0.4729	não (BH)	23
baseline_pos	pwc	cite_sources	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	23
baseline_pos	pwc	quotation_addition	gemini-3.5-flash-lite	402	1.000	no	0.9490	não (BH)	0
baseline_pos	pwc	quotation_addition	claude-haiku-4-5	402	1.000	no	0.2253	não (BH)	17
baseline_pos	pwc	quotation_addition	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	22
baseline_pos	pwc	statistics_addition	gemini-3.5-flash-lite	402	1.000	no	0.1990	não (BH)	0
baseline_pos	pwc	statistics_addition	claude-haiku-4-5	402	1.000	no	0.9151	não (BH)	23
baseline_pos	pwc	statistics_addition	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	170
baseline_pos	pwc	fluency_optimization	gemini-3.5-flash-lite	402	1.000	no	0.7860	não (BH)	0
baseline_pos	pwc	fluency_optimization	claude-haiku-4-5	402	0.0312	yes (+)	0.0039	yes (+)	15
baseline_pos	pwc	fluency_optimization	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	6
baseline_pos	pwc	easy_to_understand	gemini-3.5-flash-lite	402	0.0027	yes (−)	0.0002	yes (−)	0
baseline_pos	pwc	easy_to_understand	claude-haiku-4-5	402	0.0027	yes (−)	0.0002	yes (−)	24
baseline_pos	pwc	easy_to_understand	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	28
baseline_pos	pwc	authoritative	gemini-3.5-flash-lite	402	0.0312	yes (−)	0.0039	yes (−)	1
baseline_pos	pwc	authoritative	claude-haiku-4-5	402	0.0027	yes (−)	0.0004	yes (−)	27
baseline_pos	pwc	authoritative	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	79
baseline_pos	pwc	unique_words	gemini-3.5-flash-lite	402	0.0312	yes (−)	0.0039	yes (−)	0
baseline_pos	pwc	unique_words	claude-haiku-4-5	402	0.0027	yes (−)	0.0002	yes (−)	105
baseline_pos	pwc	unique_words	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	83
baseline_pos	pwc	technical_terms	gemini-3.5-flash-lite	402	0.3474	no	0.0549	não (BH)	0
baseline_pos	pwc	technical_terms	claude-haiku-4-5	402	0.3474	no	0.0586	não (BH)	22
baseline_pos	pwc	technical_terms	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	25
baseline_pos	pwc	keyword_stuffing	gemini-3.5-flash-lite	402	0.0027	yes (−)	0.0002	yes (−)	1
baseline_pos	pwc	keyword_stuffing	claude-haiku-4-5	402	0.0027	yes (−)	0.0002	yes (−)	61
baseline_pos	pwc	keyword_stuffing	gpt-5.6-luna	402	0.0027	yes (−)	0.0002	yes (−)	63
paired	wc	cite_sources	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	cite_sources	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	29
paired	wc	cite_sources	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	30
paired	wc	quotation_addition	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	quotation_addition	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	25
paired	wc	quotation_addition	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	26
paired	wc	statistics_addition	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	0
paired	wc	statistics_addition	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	37
paired	wc	statistics_addition	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	195
paired	wc	fluency_optimization	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	fluency_optimization	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	19
paired	wc	fluency_optimization	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	10
paired	wc	easy_to_understand	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	easy_to_understand	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	30
paired	wc	easy_to_understand	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	37
paired	wc	authoritative	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	2
paired	wc	authoritative	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	33
paired	wc	authoritative	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	95
paired	wc	unique_words	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	3
paired	wc	unique_words	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	125
paired	wc	unique_words	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	101
paired	wc	technical_terms	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	technical_terms	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	33
paired	wc	technical_terms	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	31
paired	wc	keyword_stuffing	gemini-3.5-flash-lite	500	n/d	sens.	n/d	sens.	1
paired	wc	keyword_stuffing	claude-haiku-4-5	500	n/d	sens.	n/d	sens.	78
paired	wc	keyword_stuffing	gpt-5.6-luna	500	n/d	sens.	n/d	sens.	75
baseline_pos	wc	cite_sources	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	cite_sources	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	23
baseline_pos	wc	cite_sources	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	23
baseline_pos	wc	quotation_addition	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0

continued on next page

Table 21 – continued from previous page

set	metric	technique	engine	n	p (Holm)	sig. (Holm)	p (BH)	sig. (BH)	n_citacao_zerada
baseline_pos	wc	quotation_addition	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	17
baseline_pos	wc	quotation_addition	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	22
baseline_pos	wc	statistics_addition	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	statistics_addition	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	23
baseline_pos	wc	statistics_addition	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	170
baseline_pos	wc	fluency_optimization	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	fluency_optimization	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	15
baseline_pos	wc	fluency_optimization	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	6
baseline_pos	wc	easy_to_understand	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	easy_to_understand	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	24
baseline_pos	wc	easy_to_understand	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	28
baseline_pos	wc	authoritative	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	1
baseline_pos	wc	authoritative	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	27
baseline_pos	wc	authoritative	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	79
baseline_pos	wc	unique_words	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	unique_words	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	105
baseline_pos	wc	unique_words	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	83
baseline_pos	wc	technical_terms	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	0
baseline_pos	wc	technical_terms	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	22
baseline_pos	wc	technical_terms	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	25
baseline_pos	wc	keyword_stuffing	gemini-3.5-flash-lite	402	n/d	sens.	n/d	sens.	1
baseline_pos	wc	keyword_stuffing	claude-haiku-4-5	402	n/d	sens.	n/d	sens.	61
baseline_pos	wc	keyword_stuffing	gpt-5.6-luna	402	n/d	sens.	n/d	sens.	63

Table 22: Cross-engine comparison: relative improvement (Eq. 4) of the target source per technique and per engine, over the strictly paired query set (a query enters only if it is complete on every engine and the target source occupies the same position on all of them). Reported for both query sets: `baseline_pos` (primary; queries whose target source has positive baseline visibility) and the full paired set (sensitivity) — see §3.5. **Part 5 of 5** of the complete table; the identifying columns repeat in every part.

set	metric	technique	engine	n	% zeroed
paired	pwc	cite_sources	gemini-3.5-flash-lite	500	0.2012
paired	pwc	cite_sources	claude-haiku-4-5	500	6.697
paired	pwc	cite_sources	gpt-5.6-luna	500	6.849
paired	pwc	quotation_addition	gemini-3.5-flash-lite	500	0.2012
paired	pwc	quotation_addition	claude-haiku-4-5	500	5.774
paired	pwc	quotation_addition	gpt-5.6-luna	500	5.936
paired	pwc	statistics_addition	gemini-3.5-flash-lite	500	0
paired	pwc	statistics_addition	claude-haiku-4-5	500	8.545
paired	pwc	statistics_addition	gpt-5.6-luna	500	44.52
paired	pwc	fluency_optimization	gemini-3.5-flash-lite	500	0.2012
paired	pwc	fluency_optimization	claude-haiku-4-5	500	4.388
paired	pwc	fluency_optimization	gpt-5.6-luna	500	2.283
paired	pwc	easy_to_understand	gemini-3.5-flash-lite	500	0.2012
paired	pwc	easy_to_understand	claude-haiku-4-5	500	6.928
paired	pwc	easy_to_understand	gpt-5.6-luna	500	8.447
paired	pwc	authoritative	gemini-3.5-flash-lite	500	0.4024
paired	pwc	authoritative	claude-haiku-4-5	500	7.621
paired	pwc	authoritative	gpt-5.6-luna	500	21.69
paired	pwc	unique_words	gemini-3.5-flash-lite	500	0.6036
paired	pwc	unique_words	claude-haiku-4-5	500	28.87
paired	pwc	unique_words	gpt-5.6-luna	500	23.06
paired	pwc	technical_terms	gemini-3.5-flash-lite	500	0.2012
paired	pwc	technical_terms	claude-haiku-4-5	500	7.621
paired	pwc	technical_terms	gpt-5.6-luna	500	7.078
paired	pwc	keyword_stuffing	gemini-3.5-flash-lite	500	0.2012
paired	pwc	keyword_stuffing	claude-haiku-4-5	500	18.01
paired	pwc	keyword_stuffing	gpt-5.6-luna	500	17.12
baseline_pos	pwc	cite_sources	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	cite_sources	claude-haiku-4-5	402	5.721
baseline_pos	pwc	cite_sources	gpt-5.6-luna	402	5.721
baseline_pos	pwc	quotation_addition	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	quotation_addition	claude-haiku-4-5	402	4.229
baseline_pos	pwc	quotation_addition	gpt-5.6-luna	402	5.473
baseline_pos	pwc	statistics_addition	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	statistics_addition	claude-haiku-4-5	402	5.721
baseline_pos	pwc	statistics_addition	gpt-5.6-luna	402	42.29
baseline_pos	pwc	fluency_optimization	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	fluency_optimization	claude-haiku-4-5	402	3.731
baseline_pos	pwc	fluency_optimization	gpt-5.6-luna	402	1.493
baseline_pos	pwc	easy_to_understand	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	easy_to_understand	claude-haiku-4-5	402	5.970
baseline_pos	pwc	easy_to_understand	gpt-5.6-luna	402	6.965
baseline_pos	pwc	authoritative	gemini-3.5-flash-lite	402	0.2488
baseline_pos	pwc	authoritative	claude-haiku-4-5	402	6.716
baseline_pos	pwc	authoritative	gpt-5.6-luna	402	19.65
baseline_pos	pwc	unique_words	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	unique_words	claude-haiku-4-5	402	26.12
baseline_pos	pwc	unique_words	gpt-5.6-luna	402	20.65
baseline_pos	pwc	technical_terms	gemini-3.5-flash-lite	402	0
baseline_pos	pwc	technical_terms	claude-haiku-4-5	402	5.473
baseline_pos	pwc	technical_terms	gpt-5.6-luna	402	6.219
baseline_pos	pwc	keyword_stuffing	gemini-3.5-flash-lite	402	0.2488
baseline_pos	pwc	keyword_stuffing	claude-haiku-4-5	402	15.17
baseline_pos	pwc	keyword_stuffing	gpt-5.6-luna	402	15.67
paired	wc	cite_sources	gemini-3.5-flash-lite	500	0.2012

continued on next page

Table 22 – continued from previous page

set	metric	technique	engine	n	% zeroed
paired	wc	cite_sources	claude-haiku-4-5	500	6.697
paired	wc	cite_sources	gpt-5.6-luna	500	6.849
paired	wc	quotation_addition	gemini-3.5-flash-lite	500	0.2012
paired	wc	quotation_addition	claude-haiku-4-5	500	5.774
paired	wc	quotation_addition	gpt-5.6-luna	500	5.936
paired	wc	statistics_addition	gemini-3.5-flash-lite	500	0
paired	wc	statistics_addition	claude-haiku-4-5	500	8.545
paired	wc	statistics_addition	gpt-5.6-luna	500	44.52
paired	wc	fluency_optimization	gemini-3.5-flash-lite	500	0.2012
paired	wc	fluency_optimization	claude-haiku-4-5	500	4.388
paired	wc	fluency_optimization	gpt-5.6-luna	500	2.283
paired	wc	easy_to_understand	gemini-3.5-flash-lite	500	0.2012
paired	wc	easy_to_understand	claude-haiku-4-5	500	6.928
paired	wc	easy_to_understand	gpt-5.6-luna	500	8.447
paired	wc	authoritative	gemini-3.5-flash-lite	500	0.4024
paired	wc	authoritative	claude-haiku-4-5	500	7.621
paired	wc	authoritative	gpt-5.6-luna	500	21.69
paired	wc	unique_words	gemini-3.5-flash-lite	500	0.6036
paired	wc	unique_words	claude-haiku-4-5	500	28.87
paired	wc	unique_words	gpt-5.6-luna	500	23.06
paired	wc	technical_terms	gemini-3.5-flash-lite	500	0.2012
paired	wc	technical_terms	claude-haiku-4-5	500	7.621
paired	wc	technical_terms	gpt-5.6-luna	500	7.078
paired	wc	keyword_stuffing	gemini-3.5-flash-lite	500	0.2012
paired	wc	keyword_stuffing	claude-haiku-4-5	500	18.01
paired	wc	keyword_stuffing	gpt-5.6-luna	500	17.12
baseline_pos	wc	cite_sources	gemini-3.5-flash-lite	402	0
baseline_pos	wc	cite_sources	claude-haiku-4-5	402	5.721
baseline_pos	wc	cite_sources	gpt-5.6-luna	402	5.721
baseline_pos	wc	quotation_addition	gemini-3.5-flash-lite	402	0
baseline_pos	wc	quotation_addition	claude-haiku-4-5	402	4.229
baseline_pos	wc	quotation_addition	gpt-5.6-luna	402	5.473
baseline_pos	wc	statistics_addition	gemini-3.5-flash-lite	402	0
baseline_pos	wc	statistics_addition	claude-haiku-4-5	402	5.721
baseline_pos	wc	statistics_addition	gpt-5.6-luna	402	42.29
baseline_pos	wc	fluency_optimization	gemini-3.5-flash-lite	402	0
baseline_pos	wc	fluency_optimization	claude-haiku-4-5	402	3.731
baseline_pos	wc	fluency_optimization	gpt-5.6-luna	402	1.493
baseline_pos	wc	easy_to_understand	gemini-3.5-flash-lite	402	0
baseline_pos	wc	easy_to_understand	claude-haiku-4-5	402	5.970
baseline_pos	wc	easy_to_understand	gpt-5.6-luna	402	6.965
baseline_pos	wc	authoritative	gemini-3.5-flash-lite	402	0.2488
baseline_pos	wc	authoritative	claude-haiku-4-5	402	6.716
baseline_pos	wc	authoritative	gpt-5.6-luna	402	19.65
baseline_pos	wc	unique_words	gemini-3.5-flash-lite	402	0
baseline_pos	wc	unique_words	claude-haiku-4-5	402	26.12
baseline_pos	wc	unique_words	gpt-5.6-luna	402	20.65
baseline_pos	wc	technical_terms	gemini-3.5-flash-lite	402	0
baseline_pos	wc	technical_terms	claude-haiku-4-5	402	5.473
baseline_pos	wc	technical_terms	gpt-5.6-luna	402	6.219
baseline_pos	wc	keyword_stuffing	gemini-3.5-flash-lite	402	0.2488
baseline_pos	wc	keyword_stuffing	claude-haiku-4-5	402	15.17
baseline_pos	wc	keyword_stuffing	gpt-5.6-luna	402	15.67

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, page 12, Barcelona, Spain, 2024. ACM. doi: 10.1145/3637528.3671900. arXiv:2311.09735v3.

Mirelle Bueno, Eduardo Seiti de Oliveira, Rodrigo Nogueira, Roberto A. Lotufo, and Jayr Alencar Pereira. Quati: A Brazilian Portuguese information retrieval dataset from native speakers, 2024. arXiv:2404.06976.

Conselho Administrativo de Defesa Econômica (CADE). Tribunal instaura processo administrativo contra o Google por uso de conteúdo jornalístico em resumos gerados por inteligência artificial. Julgamento do Tribunal do CADE, 23 de abril de 2026, April 2026. Processo originado de representação da Globo Comunicação e Participações S.A., com apoio de ANJ, ABERT, Ajor, ANER, FENAJ, IDEC e Artigo 19. Decisão unânime; o Tribunal caracterizou as ferramentas de opt-out existentes como "falsa escolha".

Jim Egan, Amy Ross Arguedas, Craig T. Robertson, Nic Newman, Richard Fletcher, and Rasmus Kleis Nielsen. Digital news report 2026. Technical report, Reuters Institute for the Study of Journalism, University of Oxford, 2026. URL https://reutersinstitute.politics.ox.ac.uk/

set	metric	technique	gemini med. %	haiku med. %	luna med. %
paired	pwc	cite_sources	1.740	0	-7.259
paired	pwc	quotation_addition	0.5363	0	-3.974
paired	pwc	statistics_addition	-1.324	0	-74.56
paired	pwc	fluency_optimization	0.6499	5.041	-3.927
paired	pwc	easy_to_understand	-5.255	-7.220	-27.35
paired	pwc	authoritative	-2.343	-3.846	-43.29
paired	pwc	unique_words	-2.062	-46.46	-50.02
paired	pwc	technical_terms	2.526	-0.7657	-8.906
paired	pwc	keyword_stuffing	-9.751	-25.24	-40.71
paired	wc	cite_sources	2.175	0	-7.461
paired	wc	quotation_addition	0.2370	0	-3.520
paired	wc	statistics_addition	0	0	-72.28
paired	wc	fluency_optimization	0.5522	6.660	-2.846
paired	wc	easy_to_understand	-3.068	-6.654	-26.30
paired	wc	authoritative	-2.120	-5.682	-41.28
paired	wc	unique_words	-1.665	-45.94	-48.41
paired	wc	technical_terms	2.534	-2.291	-8.004
paired	wc	keyword_stuffing	-8.423	-26.81	-39.22
baseline_pos	pwc	cite_sources	0.4816	-2.329	-12.08
baseline_pos	pwc	quotation_addition	0.0055	-2.051	-9.551
baseline_pos	pwc	statistics_addition	-1.449	0.5858	-81.53
baseline_pos	pwc	fluency_optimization	0.1682	7.886	-6.891
baseline_pos	pwc	easy_to_understand	-5.447	-10.66	-31.37
baseline_pos	pwc	authoritative	-3.132	-8.092	-48.68
baseline_pos	pwc	unique_words	-3.218	-54.12	-55.37
baseline_pos	pwc	technical_terms	1.920	-4.371	-12.41
baseline_pos	pwc	keyword_stuffing	-9.103	-28.62	-45.31
baseline_pos	wc	cite_sources	0.6922	-0.8432	-11.33
baseline_pos	wc	quotation_addition	0	-0.5784	-8.547
baseline_pos	wc	statistics_addition	-0.1738	0.7699	-80.09
baseline_pos	wc	fluency_optimization	0.3651	9.161	-6.529
baseline_pos	wc	easy_to_understand	-3.247	-10.97	-30.66
baseline_pos	wc	authoritative	-2.805	-9.352	-47.23
baseline_pos	wc	unique_words	-2.051	-52.79	-53.98
baseline_pos	wc	technical_terms	1.814	-5.965	-12.28
baseline_pos	wc	keyword_stuffing	-8.575	-30.58	-43.84

Table 23: Direction-of-effect agreement across engines, per technique: the sign of the median relative improvement on each engine and whether all engines agree. The denominator is all nine techniques of the study and an indeterminate result counts as non-agreement — the denominator is never reduced to the measurable subset. Disagreements are further split into *inversion* (opposite signs) and *attenuation* (effect present on one engine, null on another), since the two support different claims. **Part 1 of 3** of the complete table; the identifying columns repeat in every part.

set	metric	technique	gemini sign	haiku sign	luna sign	gemini CI	haiku CI
paired	pwc	cite_sources	+	0	−	yes (+)	no
paired	pwc	quotation_addition	+	0	−	no	no
paired	pwc	statistics_addition	−	0	−	no	no
paired	pwc	fluency_optimization	+	+	−	no	yes (+)
paired	pwc	easy_to_understand	−	−	−	yes (−)	yes (−)
paired	pwc	authoritative	−	−	−	yes (−)	no
paired	pwc	unique_words	−	−	−	yes (−)	yes (−)
paired	pwc	technical_terms	+	−	−	yes (+)	no
paired	pwc	keyword_stuffing	−	−	−	yes (−)	yes (−)
paired	wc	cite_sources	+	0	−	yes (+)	no
paired	wc	quotation_addition	+	0	−	no	no
paired	wc	statistics_addition	0	0	−	no	no
paired	wc	fluency_optimization	+	+	−	no	yes (+)
paired	wc	easy_to_understand	−	−	−	yes (−)	yes (−)
paired	wc	authoritative	−	−	−	no	yes (−)
paired	wc	unique_words	−	−	−	yes (−)	yes (−)
paired	wc	technical_terms	+	−	−	yes (+)	no
paired	wc	keyword_stuffing	−	−	−	yes (−)	yes (−)
baseline_pos	pwc	cite_sources	+	−	−	no	no
baseline_pos	pwc	quotation_addition	+	−	−	no	no
baseline_pos	pwc	statistics_addition	−	+	−	no	no
baseline_pos	pwc	fluency_optimization	+	+	−	no	yes (+)
baseline_pos	pwc	easy_to_understand	−	−	−	yes (−)	yes (−)
baseline_pos	pwc	authoritative	−	−	−	yes (−)	yes (−)
baseline_pos	pwc	unique_words	−	−	−	yes (−)	yes (−)
baseline_pos	pwc	technical_terms	+	−	−	yes (+)	yes (−)
baseline_pos	pwc	keyword_stuffing	−	−	−	yes (−)	yes (−)
baseline_pos	wc	cite_sources	+	−	−	no	no
baseline_pos	wc	quotation_addition	0	−	−	no	no
baseline_pos	wc	statistics_addition	−	+	−	no	no
baseline_pos	wc	fluency_optimization	+	+	−	no	yes (+)
baseline_pos	wc	easy_to_understand	−	−	−	yes (−)	yes (−)
baseline_pos	wc	authoritative	−	−	−	yes (−)	yes (−)
baseline_pos	wc	unique_words	−	−	−	yes (−)	yes (−)
baseline_pos	wc	technical_terms	+	−	−	yes (+)	yes (−)
baseline_pos	wc	keyword_stuffing	−	−	−	yes (−)	yes (−)

Table 24: Direction-of-effect agreement across engines, per technique: the sign of the median relative improvement on each engine and whether all engines agree. The denominator is all nine techniques of the study and an indeterminate result counts as non-agreement — the denominator is never reduced to the measurable subset. Disagreements are further split into *inversion* (opposite signs) and *attenuation* (effect present on one engine, null on another), since the two support different claims. **Part 2 of 3** of the complete table; the identifying columns repeat in every part.

set	metric	technique	luna CI	agree	divergence	sig. all	n_engines_sig	n eff.	n_efetivo_igual
paired	pwc	cite_sources	yes (−)	no	inversion	no	2	468	NO
paired	pwc	quotation_addition	yes (−)	no	inversion	no	1	469	NO
paired	pwc	statistics_addition	yes (−)	no	attenuation	no	1	467	NO
paired	pwc	fluency_optimization	no	no	inversion	no	1	454	NO
paired	pwc	easy_to_understand	yes (−)	yes	—	yes	3	466	NO
paired	pwc	authoritative	yes (−)	yes	—	no	2	475	NO
paired	pwc	unique_words	yes (−)	yes	—	yes	3	491	NO
paired	pwc	technical_terms	yes (−)	no	inversion	no	2	471	NO
paired	pwc	keyword_stuffing	yes (−)	yes	—	yes	3	479	NO
paired	wc	cite_sources	yes (−)	no	inversion	no	2	468	NO
paired	wc	quotation_addition	yes (−)	no	inversion	no	1	469	NO
paired	wc	statistics_addition	yes (−)	no	attenuation	no	1	467	NO
paired	wc	fluency_optimization	no	no	inversion	no	1	454	NO
paired	wc	easy_to_understand	yes (−)	yes	—	yes	3	466	NO
paired	wc	authoritative	yes (−)	yes	—	no	2	475	NO
paired	wc	unique_words	yes (−)	yes	—	yes	3	491	NO
paired	wc	technical_terms	yes (−)	no	inversion	no	2	471	NO
paired	wc	keyword_stuffing	yes (−)	yes	—	yes	3	479	NO
baseline_pos	pwc	cite_sources	yes (−)	no	inversion	no	1	402	yes
baseline_pos	pwc	quotation_addition	yes (−)	no	inversion	no	1	402	yes
baseline_pos	pwc	statistics_addition	yes (−)	no	inversion	no	1	402	yes
baseline_pos	pwc	fluency_optimization	yes (−)	no	inversion	no	2	402	yes
baseline_pos	pwc	easy_to_understand	yes (−)	yes	—	yes	3	402	yes
baseline_pos	pwc	authoritative	yes (−)	yes	—	yes	3	402	yes
baseline_pos	pwc	unique_words	yes (−)	yes	—	yes	3	402	yes
baseline_pos	pwc	technical_terms	yes (−)	no	inversion	yes	3	402	yes
baseline_pos	pwc	keyword_stuffing	yes (−)	yes	—	yes	3	402	yes
baseline_pos	wc	cite_sources	yes (−)	no	inversion	no	1	402	yes
baseline_pos	wc	quotation_addition	yes (−)	no	attenuation	no	1	402	yes
baseline_pos	wc	statistics_addition	yes (−)	no	inversion	no	1	402	yes
baseline_pos	wc	fluency_optimization	yes (−)	no	inversion	no	2	402	yes
baseline_pos	wc	easy_to_understand	yes (−)	yes	—	yes	3	402	yes
baseline_pos	wc	authoritative	yes (−)	yes	—	yes	3	402	yes
baseline_pos	wc	unique_words	yes (−)	yes	—	yes	3	402	yes
baseline_pos	wc	technical_terms	yes (−)	no	inversion	yes	3	402	yes
baseline_pos	wc	keyword_stuffing	yes (−)	yes	—	yes	3	402	yes

Table 25: Direction-of-effect agreement across engines, per technique: the sign of the median relative improvement on each engine and whether all engines agree. The denominator is all nine techniques of the study and an indeterminate result counts as non-agreement — the denominator is never reduced to the measurable subset. Disagreements are further split into *inversion* (opposite signs) and *attenuation* (effect present on one engine, null on another), since the two support different claims. **Part 3 of 3** of the complete table; the identifying columns repeat in every part.

set	metric	engine A	engine B	ρ	exact p	crit. $\lvert\rho\rvert$	$\rho \neq 0$	n tech.	excluded
paired	pwc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6781	0.0522	0.6950	no	9	—
paired	pwc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no	9	—
paired	pwc	claude-haiku-4-5	gpt-5.6-luna	0.5933	0.1010	0.6950	no	9	—
paired	wc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6781	0.0522	0.6950	no	9	—
paired	wc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no	9	—
paired	wc	claude-haiku-4-5	gpt-5.6-luna	0.5933	0.1010	0.6950	no	9	—
baseline_pos	pwc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6667	0.0589	0.7000	no	9	—
baseline_pos	pwc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5500	0.1328	0.7000	no	9	—
baseline_pos	pwc	claude-haiku-4-5	gpt-5.6-luna	0.4833	0.1938	0.7000	no	9	—
baseline_pos	wc	gemini-3.5-flash-lite	claude-haiku-4-5	0.6167	0.0857	0.7000	no	9	—
baseline_pos	wc	gemini-3.5-flash-lite	gpt-5.6-luna	0.5333	0.1475	0.7000	no	9	—
baseline_pos	wc	claude-haiku-4-5	gpt-5.6-luna	0.4833	0.1938	0.7000	no	9	—

Table 26: Pairwise Spearman rank correlation between engines over the nine-technique vector of median relative improvements. The p is exact — the permutation null over the 9! orderings is enumerated in full, since the asymptotic approximation is not valid at $n = 9$. **No pair reaches significance**: the critical $\lvert\rho\rvert$ is 0.700 (lower where ties in the ranks shrink the null), above every ρ observed here. These correlations are reported for completeness and support no claim about ordering; the per-technique effects and their intervals (Table 5) are what this study leans on.

technique	delta_comprimento_pct	gemini med. %	haiku med. %	luna med. %
cite_sources	6.725	0.4816	-2.329	-12.08
quotation_addition	3.361	0.0055	-2.051	-9.551
statistics_addition	-6.061	-1.449	0.5858	-81.53
fluency_optimization	7.310	0.1682	7.886	-6.891
easy_to_understand	-14.41	-5.447	-10.66	-31.37
authoritative	-13.51	-3.132	-8.092	-48.68
unique_words	-9.353	-3.218	-54.12	-55.37
technical_terms	-0.7979	1.920	-4.371	-12.41
keyword_stuffing	-12.43	-9.103	-28.62	-45.31

Table 27: Length change of the transformed source per technique (median over the 525 queries, in words, against the original source), alongside the median visibility change each engine measures. The transformation is performed once and reused across engines, so the length column is engine-independent. Across the nine techniques, length change and visibility change are strongly rank-correlated: $\rho = +0.80$ ($p = 0.014$) on gemini, $+0.73$ ($p = 0.031$) on haiku and $+0.67$ ($p = 0.059$) on luna, exact permutation p with critical $\lvert\rho\rvert = 0.700$. See §5.7 for what this does and does not confound.

technique	engine	n_respostas	n_abstencao	pct_abstencao	queda_vs_baseline_pp
baseline	gemini-3.5-flash-lite	75	75	100.0	n/d
baseline	claude-haiku-4-5	75	75	100.0	n/d
baseline	gpt-5.6-luna	75	75	100.0	n/d
cite_sources	gemini-3.5-flash-lite	75	75	100.0	0
cite_sources	claude-haiku-4-5	75	75	100.0	0
cite_sources	gpt-5.6-luna	75	75	100.0	0
quotation_addition	gemini-3.5-flash-lite	75	70	93.33	6.667
quotation_addition	claude-haiku-4-5	75	72	96.00	4.000
quotation_addition	gpt-5.6-luna	75	72	96.00	4.000
statistics_addition	gemini-3.5-flash-lite	75	51	68.00	32.00
statistics_addition	claude-haiku-4-5	75	60	80.00	20.00
statistics_addition	gpt-5.6-luna	75	56	74.67	25.33
fluency_optimization	gemini-3.5-flash-lite	75	74	98.67	1.333
fluency_optimization	claude-haiku-4-5	75	75	100.0	0
fluency_optimization	gpt-5.6-luna	75	74	98.67	1.333
easy_to_understand	gemini-3.5-flash-lite	75	64	85.33	14.67
easy_to_understand	claude-haiku-4-5	75	72	96.00	4.000
easy_to_understand	gpt-5.6-luna	75	72	96.00	4.000
authoritative	gemini-3.5-flash-lite	75	63	84.00	16.00
authoritative	claude-haiku-4-5	75	66	88.00	12.00
authoritative	gpt-5.6-luna	75	69	92.00	8.000
unique_words	gemini-3.5-flash-lite	75	75	100.0	0
unique_words	claude-haiku-4-5	75	75	100.0	0
unique_words	gpt-5.6-luna	75	75	100.0	0
technical_terms	gemini-3.5-flash-lite	75	71	94.67	5.333
technical_terms	claude-haiku-4-5	75	75	100.0	0
technical_terms	gpt-5.6-luna	75	72	96.00	4.000
keyword_stuffing	gemini-3.5-flash-lite	75	62	82.67	17.33
keyword_stuffing	claude-haiku-4-5	75	67	89.33	10.67
keyword_stuffing	gpt-5.6-luna	75	69	92.00	8.000

Table 28: Abstention on the 25 pitfall queries, by technique and engine: the share of generated answers in which the engine states that the provided sources do not answer the question. At baseline all three engines abstain in 100% of answers; every figure below that is erosion of a correct refusal. Detection is by lexical pattern over PT-BR refusal phrasings, hand-checked on a sample — not a validated classifier (§6); it is reliable here only because the baseline refusal is a near-verbatim fixed phrase.

technique	metric	n_pegadinhas	baseline_total_mean	baseline_ci_lo	baseline_ci_hi	tecnica_total_mean	tecnica_ci_lo	tecnica_ci_hi	delta_mean	delta_ci_lo	delta_ci_hi
cite_sources	pwc	25	0	0	0	0.0533	0	0.1333	0.0533	0	0.1333
cite_sources	wc	25	0	0	0	0.0533	0	0.1333	0.0533	0	0.1333
quotation_addition	pwc	25	0	0	0	0.1972	0.0620	0.3544	0.1972	0.0620	0.3544
quotation_addition	wc	25	0	0	0	0.2271	0.0750	0.3956	0.2271	0.0750	0.3956
statistics_addition	pwc	25	0	0	0	0.3150	0.1578	0.4798	0.3150	0.1578	0.4798
statistics_addition	wc	25	0	0	0	0.3724	0.1932	0.5560	0.3724	0.1932	0.5560
fluency_optimization	pwc	25	0	0	0	0.0385	0	0.1037	0.0385	0	0.1037
fluency_optimization	wc	25	0	0	0	0.0449	0	0.1165	0.0449	0	0.1165
easy_to_understand	pwc	25	0	0	0	0.2079	0.0896	0.3456	0.2079	0.0896	0.3456
easy_to_understand	wc	25	0	0	0	0.2469	0.1067	0.4069	0.2469	0.1067	0.4069
authoritative	pwc	25	0	0	0	0.2140	0.0765	0.3768	0.2140	0.0765	0.3768
authoritative	wc	25	0	0	0	0.2335	0.0800	0.4000	0.2335	0.0800	0.4000
unique_words	pwc	25	0	0	0	0	0	0	0	0	0
unique_words	wc	25	0	0	0	0	0	0	0	0	0
technical_terms	pwc	25	0	0	0	0.0537	0	0.1366	0.0537	0	0.1366
technical_terms	wc	25	0	0	0	0.0667	0	0.1733	0.0667	0	0.1733
keyword_stuffing	pwc	25	0	0	0	0.3148	0.1524	0.4851	0.3148	0.1524	0.4851
keyword_stuffing	wc	25	0	0	0	0.3510	0.1771	0.5301	0.3510	0.1771	0.5301

Table 29: Citation induction on pitfall queries (unanswerable-by-design): total Imp_{wc}/Imp_{pwc} across all 5 sources, by technique. A value significantly above zero flags a technique that induced the engine to cite a source that does not answer the query. **Part 1 of 2** of the complete table; the identifying columns repeat in every part.

technique	metric	n_pegadinhas	inducao
cite_sources	pwc	25	no
cite_sources	wc	25	no
quotation_addition	pwc	25	yes
quotation_addition	wc	25	yes
statistics_addition	pwc	25	yes
statistics_addition	wc	25	yes
fluency_optimization	pwc	25	no
fluency_optimization	wc	25	no
easy_to_understand	pwc	25	yes
easy_to_understand	wc	25	yes
authoritative	pwc	25	yes
authoritative	wc	25	yes
unique_words	pwc	25	no
unique_words	wc	25	no
technical_terms	pwc	25	no
technical_terms	wc	25	no
keyword_stuffing	pwc	25	yes
keyword_stuffing	wc	25	yes

Table 30: Citation induction on pitfall queries (unanswerable-by-design): total Imp_{wc}/Imp_{pwc} across all 5 sources, by technique. A value significantly above zero flags a technique that induced the engine to cite a source that does not answer the query. **Part 2 of 2** of the complete table; the identifying columns repeat in every part.

set	technique	paper PAWC	paper Δ%	our median %	CI lo	CI hi	nosso_melhoria_media_pct
baseline_pos	cite_sources	24.60	27.46	1.809	0.2001	3.904	15.35
baseline_pos	quotation_addition	27.20	40.93	0.5429	-0.6744	2.595	11.77
baseline_pos	statistics_addition	25.20	30.57	-1.347	-3.286	0.7224	11.02
baseline_pos	fluency_optimization	24.70	27.98	0.6723	-0.3636	2.483	10.75
baseline_pos	easy_to_understand	22.00	13.99	-5.365	-7.573	-3.571	5.697
baseline_pos	authoritative	21.30	10.36	-2.358	-4.307	-0.2585	8.970
baseline_pos	unique_words	20.50	6.218	-2.139	-4.166	-0.2178	1.817
baseline_pos	technical_terms	22.70	17.62	2.553	0.9487	4.282	11.73
baseline_pos	keyword_stuffing	17.70	-8.290	-9.843	-11.70	-5.697	-4.019
all	cite_sources	24.60	27.46	1.740	0.2001	3.796	15.32
all	quotation_addition	27.20	40.93	0.5363	-0.6521	2.132	11.75
all	statistics_addition	25.20	30.57	-1.324	-3.114	0.5866	10.98
all	fluency_optimization	24.70	27.98	0.6499	-0.3601	2.313	10.71
all	easy_to_understand	22.00	13.99	-5.255	-7.573	-3.459	5.686
all	authoritative	21.30	10.36	-2.343	-4.193	0	8.934
all	unique_words	20.50	6.218	-2.062	-4.230	-0.2078	1.810
all	technical_terms	22.70	17.62	2.526	0.9487	4.084	11.70
all	keyword_stuffing	17.70	-8.290	-9.751	-11.66	-5.590	-4.003

Table 31: Direction-of-effect comparison against the original paper's Table 1 (PAWC-Overall, Aggarwal et al. 2024) and Spearman rank correlation. Absolute scales are not comparable across studies (different engines/metrics normalization) — only sign and ordering are. Reported for both query sets: the Spearman ρ differs between them because medians pinned at exactly 0% in the full set become ties and receive averaged ranks; `baseline_pos` is primary. **Part 1 of 2** of the complete table; the identifying columns repeat in every part.

`digital-news-report/2026/brazil`. Brazil country page: weekly use of AI chatbots for news at 13%, against a 10% global average.

Olivier Martinez. Optimizing visibility in generative engines: A critical survey of generative engine optimization (2023–2026), 2026. arXiv:2607.14035, submitted 15 July 2026; critical survey of 45 GEO studies.

Haritz Puerto, Martin Gubri, Tommaso Green, Seong Joon Oh, and Sangdoo Yun. C-SEO bench: Does conversational SEO work? In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025. arXiv:2506.11097.

Tardelli Ronan Coelho Stekel. MTEB-BR: A text embedding benchmark for Brazilian Portuguese, 2026. arXiv:2607.04581; 22 native PT-BR tasks, translations excluded by construction; v1 circulated as MTEB-PT.

Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. AutoGEO; arXiv:2510.11438.

Jiaqi Yuan, Jialu Wang, Zihan Wang, Qingyun Sun, Ruijie Wang, and Jianxin Li. AgenticGEO: A self-evolving agentic system for generative engine optimization, 2026. arXiv:2603.20213, submitted 2 March 2026.

set	technique	n	CI excl. 0	replicates
baseline_pos	cite_sources	497	yes (+)	yes
baseline_pos	quotation_addition	497	no	yes
baseline_pos	statistics_addition	497	no	no
baseline_pos	fluency_optimization	497	no	yes
baseline_pos	easy_to_understand	497	yes (−)	no
baseline_pos	authoritative	497	yes (−)	no
baseline_pos	unique_words	497	yes (−)	no
baseline_pos	technical_terms	497	yes (+)	yes
baseline_pos	keyword_stuffing	497	yes (−)	yes
all	cite_sources	500	yes (+)	yes
all	quotation_addition	500	no	yes
all	statistics_addition	500	no	no
all	fluency_optimization	500	no	yes
all	easy_to_understand	500	yes (−)	no
all	authoritative	500	yes (−)	no
all	unique_words	500	yes (−)	no
all	technical_terms	500	yes (+)	yes
all	keyword_stuffing	500	yes (−)	yes

Table 32: Direction-of-effect comparison against the original paper's Table 1 (PAWC-Overall, Aggarwal et al. 2024) and Spearman rank correlation. Absolute scales are not comparable across studies (different engines/metrics normalization) — only sign and ordering are. Reported for both query sets: the Spearman ρ differs between them because medians pinned at exactly 0% in the full set become ties and receive averaged ranks; `baseline_pos` is primary. **Part 2 of 2** of the complete table; the identifying columns repeat in every part.

categoria	n_chamadas_engine	tokens_input_medio	tokens_output_medio	custo_usd_total_engine	custo_usd_medio_chamada	n_chamadas_transform	tokens_input_medio_transform
baseline	1575	2713.1	312.0	1.275	0.0008	0	n/d
cite_sources	1575	2727.4	312.0	1.279	0.0008	525	636.8
quotation_addition	1575	2712.2	309.6	1.269	0.0008	525	626.8
statistics_addition	1575	2714.0	314.1	1.280	0.0008	525	637.8
fluency_optimization	1575	2728.8	311.0	1.277	0.0008	525	629.8
easy_to_understand	1575	2595.4	311.2	1.245	0.0008	525	633.8
authoritative	1575	2654.9	309.1	1.255	0.0008	525	630.8
unique_words	1575	2751.4	313.9	1.289	0.0008	525	628.8
technical_terms	1575	2758.3	317.6	1.297	0.0008	525	621.8
keyword_stuffing	1575	2629.9	304.6	1.241	0.0008	525	630.8
TOTAL	15750	n/d	n/d	n/d	n/d	4725	n/d

Table 33: API cost table: tokens and US$ per technique, aggregated directly from execution traces. **Part 1 of 2** of the complete table; the identifying columns repeat in every part.

tokens_output_medio_transform	custo_usd_total_transform	custo_usd_total
n/d	0	1.275
490.8	0.3788	1.657
475.5	0.3682	1.638
477.4	0.3704	1.651
492.1	0.3793	1.657
358.8	0.2907	1.536
418.2	0.3303	1.586
515.2	0.3944	1.683
521.7	0.3983	1.695
393.3	0.3136	1.555
n/d	n/d	15.93

Table 34: API cost table: tokens and US$ per technique, aggregated directly from execution traces. **Part 2 of 2** of the complete table; the identifying columns repeat in every part.