Phantm.

INDEPENDENT EVALUATION

PHANTM EVALUATION REPORT

47.1% cost reduction across 13,491 LLM requests.

A two-phase evaluation of an adaptive LLM optimization pipeline on general AI benchmarks and production customer-experience workloads.

> Across **13,491 prompts** spanning nine general AI benchmarks and six production CX workloads, Phantm reduced inference cost by **47.1%** against a realistic enterprise baseline. The deterministic-benchmark accuracy gap is **0.013 ± 0.019**; every LLM-judged source passes TOST equivalence at **±0.2** on a 5-point Likert scale.

Phantm
OpenAI-compatible LLM optimization proxy
getphantm.ai

May 2026
Independent Evaluation

n = 13,491 · TOST ±0.2 ✓ · P95 206ms

Contents

§ 1 — SUMMARY

The headline result, and what's behind it.

Phantm is a drop-in LLM optimization proxy. It exposes an OpenAI-compatible endpoint and applies a six-stage adaptive pipeline to every request before forwarding it to the upstream provider. No SDK changes, no prompt rewrites, no model swapping by the customer.

This report covers a **13,491-prompt** evaluation conducted across two phases: a general AI benchmark phase (6,500 prompts across nine standard benchmarks) and a customer-experience production workload phase (6,991 prompts across six CX datasets spanning retail, banking, ecommerce, multi-domain dialog, and agentic vertical support). Across the full evaluation, Phantm reduced inference cost by **47.1%** against a realistic enterprise baseline. Quality was preserved: the deterministic-benchmark accuracy gap is `0.013 ± 0.019`, and every LLM-judged source shows a quality gap below the threshold at which users notice a difference, per established rating-scale research (see § 3.5).

COST REDUCTION · COMBINED

47.1%

n = 13,491 · $57.97 → $30.68

CX PRODUCTION WORKLOADS

60.1%

n = 6,991 · $27.52 → $10.97

END-TO-END OVERHEAD

<200ms

P50 181.6 · P95 206.0

1.1 Cost

TABLE 1 — INFERENCE COST BY PHASE

PHASE	BASELINE	OPTIMIZED	SAVINGS	REDUCTION
General benchmarks	$30.46	$19.71	$10.75	35.3%
CX production workloads	$27.52	$10.97	$16.55	60.1%
Combined	**$57.97**	**$30.68**	**$27.30**	47.1%

1.2 Quality

TABLE 2 — QUALITY GAPS, GENERAL VS. CX

METRIC	GENERAL	CX
Deterministic benchmark accuracy gap (aggregate)	**0.013** ± 0.019	—
LLM-judge quality gap (overall, 5-pt scale)	**0.091** ± 0.039	**0.067** ± 0.020

All CX responses were LLM-judged; deterministic graders apply only to the general benchmark phase.

The remainder of this report describes what Phantm does, how the evaluation was constructed, and the full per-source results.

§ 2 — WHAT PHANTM DOES

Six adaptive stages on every request.

Phantm sits between an application and its LLM providers as an OpenAI-compatible proxy. Every request passes through an adaptive optimization pipeline that runs in real time before the upstream call is dispatched.

The pipeline is composed of six stages, each of which examines the request and decides whether and how to act based on its content. Total pipeline overhead is under **200 milliseconds**, and no application-side changes are required to use it.

FIGURE 1 · PIPELINE ARCHITECTURE

Six adaptive stages from incoming API call to upstream provider response.

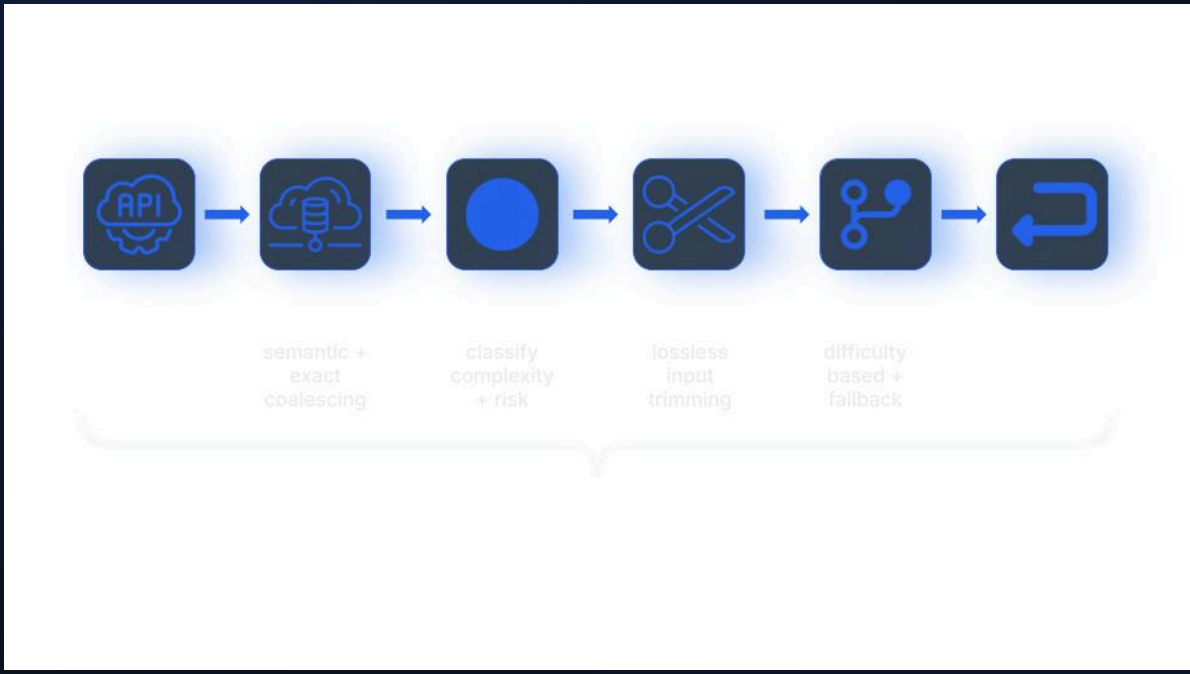

Figure 1. Pipeline architecture. Every request enters at the **Application** endpoint, traverses six adaptive stages, and is dispatched to the upstream **Provider**. **Pruning** is an escape-hatch stage that fires only when an input risks context-overflow (activation < 0.1% in this evaluation).

The differentiation is in the integration: every request passing through Phantm has all six stages applied adaptively, with no application-side changes. The cost reductions reported in this evaluation are the joint result of these stages firing across 13,491 requests — not the contribution of any single stage.

§ 3 — EVALUATION DESIGN

Two phases. Public benchmarks plus production workloads.

3.1 Two-phase design

The evaluation was structured in two phases to answer two distinct questions. The general benchmark phase asks whether the optimization pipeline preserves quality on tasks where quality is measurable against known-correct answers. The CX phase asks how much the pipeline saves on workloads that resemble actual customer-facing deployment traffic — long, repeated system prompts; many short user turns per prompt; conversation histories of varying length. Both phases share the same optimization configuration and the same baseline assignment logic.

3.2 Corpus

General benchmarks (6,500 prompts)

TABLE 3 — GENERAL BENCHMARK CORPUS

SOURCE	N	TASK TYPE	GRADING
WildChat	1,500	Open-domain chat	LLM-judged contrastive
HotPotQA	1,000	Factoid QA	LLM-graded EM/F1
MMLU	1,000	Multiple choice QA	Deterministic accuracy
BFCL	750	Tool use	Deterministic tool-call equivalence
BBH	500	Reasoning	LLM-graded exact match
LongBench	500	Long-context QA	LLM-judged contrastive
Hermes FC	500	Tool use (chat)	LLM-judged tool use
IFEval	500	Instruction following	Deterministic constraint validation
DialogSum	250	Summarization	LLM-judged contrastive

CX production workloads (6,991 prompts across 21 distinct system prompts)

TABLE 4 — CX PRODUCTION CORPUS

SOURCE	N	CX VERTICAL	TOOL USE
ABCD	1,750	Retail customer support	No
Nemotron	1,500	Agentic vertical CS (rentals, parking, security, sports retail, theme park, vet telehealth)	Yes
Taskmaster-2	1,500	Multi-vertical (food, hotels, movies, sports, flights, restaurant search)	No
Banking77	750	Banking	No
Bitext	750	Ecommerce support	No
MultiWOZ	741	Multi-domain (attraction, hotel, restaurant, taxi)	No

The CX corpus is structured around 21 distinct system prompts averaging 333 user queries per prompt (median 250, range 179 to 1,750). This mirrors real CX traffic, where a small number of stable prompts are reused across high volumes of customer interaction. This structure also allows the provider cache features to fire naturally rather than being artificially simulated.

3.3 Baseline assignment

Each prompt in the corpus is baselined against a specific model chosen to represent how that workload would be deployed without Phantm. Two model tiers are used in the baseline: `solver tier` (`gpt-5.4` / `claude-sonnet-4-6`) and `mini tier` (`gpt-5.4-mini` / `claude-haiku-4-5`). Baseline tier is assigned per source based on the technical complexity of the task: long-context, tool-use, and reasoning-heavy sources are baselined to solver tier (matching the standard production deployment for these task types in enterprise settings); open-domain chat, factoid QA, multiple-choice QA, summarization, and instruction-following sources are baselined to mini tier. A token-floor rule promotes any general-benchmark request with input length above 3,000 tokens to solver tier regardless of source, to avoid forcing a mid-tier model on inputs that exceed its effective working memory.

Provider track is assigned stratified-randomly per source with a target split of 60% OpenAI / 40% Anthropic.

3.4 Scoring

General benchmarks are scored using a mix of deterministic graders and LLM-correctness judges. MMLU, IFEval, and BFCL are scored fully deterministically (multiple-choice match, constraint validation, tool-call equivalence). HotPotQA and BBH are scored by an LLM-correctness judge that confirms whether the extracted answer matches the reference; the agreement rate between deterministic and LLM-correctness scoring on overlapping MMLU rows was **88.1%**. Sources without ground-truth answers (WildChat, LongBench, DialogSum, Hermes FC) are scored by an LLM judge that compares the optimized and baseline responses directly.

CX workloads are fully LLM-judged. Two rubrics are used: a general contrastive rubric for conversational sources (5,491 rows) and a tool-use rubric for the agentic CS corpus (1,500 rows on `nemotron`).

Cross-provider judging is used throughout. OpenAI-track responses are judged by `claude-opus-4-6`. Anthropic-track responses are judged by `gpt-5.4`. This eliminates the self-preference effect that occurs when a model judges its own outputs. The judge scores each response on a 1–5 scale across multiple quality dimensions. The A/B position of the two responses presented to the judge is deterministically flipped per record using a SHA-256 hash of the record ID, mitigating position bias.

*For clarity throughout the rest of this report, **cache hit rate** refers to the fraction of input tokens served from cache, computed from row-level `cache_read` and `cached_input` token counts against total input tokens.*

3.5 Equivalence testing

> We test whether Phantm's quality matches the baseline using the standard two one-sided tests (TOST) procedure (Lakens, 2017) against a threshold of **`±0.2 points`** on the 5-point Likert scale. A source passes equivalence at this threshold if its entire confidence interval falls within ±0.2.

The `0.2` threshold draws from two well-established results. Research on patient-reported outcomes (Norman, Sloan & Wyrwich, *Medical Care*, 2003) finds that users typically can't reliably perceive differences smaller than about half a standard deviation — roughly `0.5` points on a 5-point Likert scale. Studies that ask raters directly to identify the smallest meaningful difference (Anvari & Lakens, *Journal of Experimental Social Psychology*, 2021) put the threshold even lower, in the `0.20–0.39` range for single Likert items. We use `0.2` — the strictest threshold supported by this literature — as the equivalence band.

The standard deviation of per-row quality deltas in this evaluation is approximately `0.83` across all LLM-judged rows (and `0.90` for the general phase considered separately) — materially below the per-rating noise floor of 1.0–1.2 documented in the LLM-as-judge literature, reflecting the variance reduction from paired contrastive judging and dimension averaging. These empirical SDs are used

to compute the confidence intervals reported throughout the rest of this report. For deterministic-benchmark accuracy, the confidence interval is computed from row-level paired binary outcomes.

3.6 Relation to prior work

Public evaluations of LLM cost-optimization products fall into two patterns. Academic frameworks — most notably Martian's RouterBench (Hu et al., 2024) and LMSYS's RouteLLM (Ong et al., 2024) — publish full methodology, datasets, and code, summarizing cost-quality tradeoffs through point comparisons or composite scalars. Vendor reports — including those from Not Diamond, Portkey, and several gateway products — typically lead with headline cost reductions and report response quality either informally or not at all.

To our knowledge, **no prior public evaluation in this product category has used formal equivalence testing against a pre-set threshold.** The methodology — TOST against a smallest-effect-size-of-interest drawn from measurement-theory literature — is the standard procedure for testing equivalence claims in clinical research and the social sciences (Lakens, 2017; Norman et al., 2003).

§ 4 — RESULTS

Cost, savings, quality, latency.

4.1 Cost

TABLE 5 — INFERENCE COST (USD) BY PHASE

PHASE	BASELINE	OPTIMIZED	SAVINGS	REDUCTION
General benchmarks	$30.46	$19.71	$10.75	35.3%
CX production workloads	$27.52	$10.97	$16.55	60.1%
Combined	**$57.97**	**$30.68**	**$27.30**	47.1%

CX savings are substantially higher than general benchmark savings for two reasons. First, the CX corpus structure (stable system prompts, many queries per prompt) naturally enables cache orchestration to amortize over many requests. Second, CX traffic contains a larger fraction of straightforward requests that route safely to smaller models than the general benchmark mix does. Both of these conditions resemble real production CX deployments more closely than the general benchmark corpus does.

The reported cost reductions are **net of all routing decisions in both directions**. While the optimizer routes most requests to smaller models (§ 4.2), it also up-routes a small fraction to more capable models when the gate classifier identifies the input as too complex for the baseline tier. Approximately 1.5% of general-benchmark requests and 0.2% of CX requests were up-routed in this way, typically to a top-tier reasoning model. These up-routes increase per-request cost on the affected rows but preserve response quality where the baseline would have been insufficient — they are a feature of the routing logic, not an exception to the cost story.

4.2 Where the savings come from

Phantm's optimization pipeline applies six stages adaptively to each request. The cost reduction is the joint result of these stages, not the contribution of any single one. The activation rates below indicate, for each stage, the fraction of requests on which the stage did meaningful work.

TABLE 6 — STAGE ACTIVATION RATES

STAGE	GENERAL ACTIVATION	CX ACTIVATION	NOTES
Adaptive model routing	60.8%	72.8%	Fraction of requests routed away from baseline tier
Output shaping	94.6%	75.1%	Length and format guidance applied to most requests
Semantic compression	22.1%	45.1%	Fires when conversation history is present and compressible
Prompt cleanup	14.0%	23.5%	Fires when input contains compressible filler
Provider cache (Anthropic / OpenAI input-token hit rate)	—	89.8% / 36.3%	Cache hit rates on each provider track
Pruning	0.2%	<0.1%	Reserved for context-overflow risk; negligible in benchmark traffic

4.3 Quality

Quality is reported across two views: deterministic benchmark accuracy (the most concrete measure of correctness) and quality gaps from LLM judging (the closest signal to user-perceptible quality across heterogeneous tasks). Both views address different questions and should be read together. Confidence intervals and equivalence tests use the methodology described in § 3.5.

> **A note on win/tie/loss verdicts.** Contrastive forced-choice judging — the industry-standard metric for LLM evaluation — requires the judge to pick a winner even when both responses are functionally equivalent, which artificially amplifies small underlying differences. Raw score gaps and deterministic accuracy, with their associated confidence intervals and equivalence tests, are the more representative measures of user-perceptible quality.

Deterministic benchmark accuracy

TABLE 7 — DETERMINISTIC BENCHMARK ACCURACY

BENCHMARK	GRADER	N	BASELINE	OPTIMIZED	GAP
HotPotQA	Normalized EM/F1	1,000	0.867	0.846	0.021
MMLU	Multiple-choice match	1,000	0.777	0.760	0.017
IFEval	Constraint validation	470	0.706	0.696	0.011
BFCL	Tool-call equivalence	723	0.588	0.606	0
BBH	Normalized exact match	500	0.850	0.818	0.032

Aggregated across all five graders, n-weighted: **gap = 0.013 ± 0.019**. The interval crosses zero — the data are consistent with no real difference. Accuracy on objectively-graded tasks is essentially flat. Three of five benchmarks have gaps under 2 percentage points; tool-use accuracy (BFCL) is functionally equivalent; the largest gap is on BBH at 3.2 percentage points. The aggregate gap of 0.013 on a 0–1 scale is below what would be considered a meaningful regression on any of these benchmarks individually.

LLM-judge quality gaps

TABLE 8 — PER-DIMENSION QUALITY GAPS (5-POINT LIKERT SCALE)

DIMENSION	GENERAL	CX
Accuracy	0.034	0.025
Clarity	0.096	0.091
Helpfulness	0.144	0.085
Overall	**0.091**	**0.067**

Overall quality gaps are **0.091 ± 0.039** in the general phase and **0.067 ± 0.020** in the CX phase, on a 5-point scale. Both averages sit well below the ±0.2 equivalence threshold (§ 3.5) — the general phase at less than half the threshold, the CX phase at roughly one-third. Both are below the level at which users would notice a difference. The accuracy dimension is the most concrete measure of correctness within the LLM-judge framework and shows the smallest gap in both phases (0.025 and 0.034).

CX per-source quality (TOST equivalence at ±0.2)

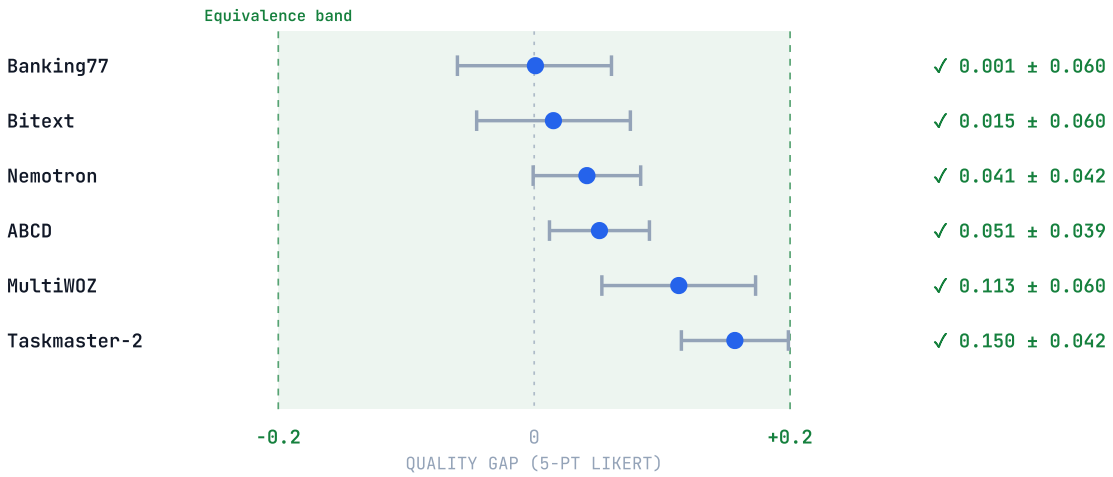

Figure 2. Per-source quality-gap confidence intervals against the ±0.2 equivalence band. Every CX source passes TOST equivalence.

TABLE 9 — CX PER-SOURCE QUALITY GAP

SOURCE	VERTICAL	N	QUALITY GAP	TOST AT ±0.2
Banking77	Banking	750	0.001 ± 0.060	✓ pass
Bitext	Ecommerce CS	750	0.015 ± 0.060	✓ pass
Nemotron	Agentic CS	1,497	0.041 ± 0.042	✓ pass
ABCD	Retail support	1,750	0.051 ± 0.039	✓ pass
MultiWOZ	Multi-domain	741	0.113 ± 0.060	✓ pass
Taskmaster-2	Multi-vertical	1,500	0.150 ± 0.042	✓ pass

All six CX sources have quality gaps that pass equivalence at the ±0.2 threshold (§ 3.5) — meaning each gap, even at the upper end of its interval, stays below the level at which users would notice a difference. Three sources (Banking77, Bitext, Nemotron) have intervals that cross zero, meaning the gap cannot be statistically distinguished from no difference at all.

4.4 Latency

End-to-end pipeline overhead, warm CX state:

TABLE 10 — END-TO-END OVERHEAD

METRIC	END-TO-END (MS)	OPTIMIZER-INTERNAL (MS)
Mean	178.1	153.2
Median (P50)	181.6	157.6
P95	206.0	178.5

TABLE 11 — PER-STAGE BREAKDOWN

STAGE	MEAN MS
Gate (real-time classifier)	130.7
Compression	13.2
Routing	4.1
Provider cache lookup / padding	2.4
Cleanup	1.2
Pruning	0.5

The real-time classifier that drives adaptive routing dominates the latency budget. All other stages combined add under **22 milliseconds**. At ~170ms end-to-end, the pipeline adds less than 5% to the latency of a typical upstream LLM call — well below any threshold at which users would notice an additional delay.

§ 5 — CONCLUSION

What this means in production.

Across **13,491 prompts** spanning general AI benchmarks and production customer-experience workloads, Phantm reduced inference cost by **47.1%** with no compromise to response quality.

Accuracy on deterministic benchmarks was effectively unchanged from the baseline. Every LLM-judged source produced responses whose quality gap fell below the threshold at which users would notice a difference, and three of the six CX sources showed no measurable difference from the baseline at all.

Customers can deploy Phantm in front of existing LLM workloads to capture these savings without changing their application code, swapping models, or sacrificing the quality of the responses their users receive.

§ 6 — LIMITATIONS

Where this evaluation doesn't reach.

1. **Semantic cache activation is low in this evaluation.** Semantic caching fired on 36 of 6,991 CX requests (0.51%). This rate reflects the diversity of the benchmark corpus — 21 distinct system prompts across seven verticals with limited naturally-occurring query repetition. Production deployments with single-customer traffic patterns, where the same questions recur frequently from the same underlying user base, will show substantially higher semantic cache hit rates.
2. **Benchmark corpus is public.** The general benchmark corpus is drawn from widely-used public datasets. Real customer traffic differs from these distributions in prompt length, query complexity, language, and domain mix. The CX corpus is closer to production traffic in structure but is still benchmark-derived. Production results in any specific deployment will depend on the workload's particular characteristics.
3. **Single optimizer configuration.** This evaluation runs a single set of optimizer settings against the baseline. The configuration is tuned for general-purpose CX deployment; customers with specialized workloads (e.g., heavy creative generation, code completion, long-form reasoning) may benefit from adjusted thresholds and feature weights.

APPENDIX A · ROUTING DISTRIBUTION

From → To, by phase.

Aggregate routing transitions for each of the two evaluation phases. Downroutes (request handled by a smaller-tier model than the baseline) are the majority of decisions in both phases.

TABLE 12 — CX PHASE ROUTING (N = 6,991)

FROM → TO	N	SHARE
Full → mini downroute	1,754	25.1%
Full → nano downroute	1,535	22.0%
Mini → mini same	1,587	22.7%
Mini → nano downroute	1,495	21.4%
Nano → nano same	310	4.4%
Nano → mini uproute	288	4.1%
Mini → full uproute	14	0.2%
Full → full same	6	0.1%
Nano → full uproute	2	0.0%
Aggregate downroute	4,784	68.4%

TABLE 13 — GENERAL BENCHMARKS ROUTING (N = 6,234)

FROM → TO	N	SHARE
Mini → nano downroute	1,995	32.0%
Mini → mini same	1,276	20.5%
Nano → nano same	1,115	17.9%
Full → mini downroute	1,022	16.4%
Full → nano downroute	507	8.1%
Nano → mini uproute	204	3.3%
Full → full same	55	0.9%
Mini → full uproute	54	0.9%
Nano → full uproute	6	0.1%
Aggregate downroute	3,524	56.5%

REFERENCES

Cited work.

[1] Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, 96, 104159.

[2] Hu, Q. J., et al. (2024). RouterBench: A Benchmark for Multi-LLM Routing System. arXiv:2403.12031.

[3] Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.

[4] Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, 41(5), 582–592.

[5] Ong, I., et al. (2024). RouteLLM: Learning to Route LLMs with Preference Data. arXiv:2406.18665.