TypeSafe Jev vs. Frontier Claude on Multiple-Choice Benchmarks

A reproducible, contamination-aware comparison

TypeSafe evaluation suite

September 18, 2026

Abstract

We compare the TypeSafe System One `Choice` primitive (Jev) against the current frontier Claude models—Haiku 4.5, Sonnet 5, and Opus 5—on four multiple-choice benchmarks: RACE-H (passage-grounded reading comprehension), CommonsenseQA (commonsense knowledge), MMLU-CF (a contamination-free MMLU variant), and Grace-Coolidge (a hand-authored, contamination-controlled corpus over a single long article). Each Claude model is run at three reasoning settings—*none* (extended thinking disabled, output restricted to the answer letter), *low*, and *medium*—so the accuracy/cost trade-off of reasoning is visible directly. Among the no-reasoning, answer-only configurations TypeSafe Jev is the top measured system on CommonsenseQA and MMLU-CF and ties Opus 5 on RACE-H. Enabling reasoning lets Opus 5 overtake Jev by one to five points on the public benchmarks—but at two to twelve times the cost and roughly ten times the latency, and with no benefit on Grace-Coolidge, where Opus even dips. Reasoning's clearest win is repairing the truncation penalty that the no-reasoning constraint inflicts on Haiku 4.5. Across settings Jev remains the best accuracy-per-dollar and accuracy-per-latency system, and on long-context items it is over *two orders of magnitude cheaper* (0.84 vs 129.69 USD per 1,000 questions).

1 Motivation

Public leaderboard numbers for reading-comprehension and knowledge benchmarks are both *stale* (they predate the current model generation) and potentially *contaminated* (benchmark items may sit in pretraining data). Rather than trust historical figures, we re-run the current models ourselves under a single fixed protocol, and we include MMLU-CF—an explicitly contamination-free benchmark—as a control. All numbers in this paper are regenerated by the scripts in this repository from a fixed random seed.

2 Setup

2.1 Datasets

- **RACE-H** (`ehovy/race`, high-school split): passage + question, four options; answers are grounded in the passage.
- **CommonsenseQA** (`tau/commonsense_qa`, dev split): question only, five options; requires world knowledge.

- **MMLU-CF** (`microsoft/MMLU-CF`, val split): multi-domain knowledge questions, four options; a contamination-free MMLU variant.

For these three public benchmarks we sample $n = 100$ items with seed 42; every system sees the identical items. A fourth, hand-authored benchmark (Grace-Coolidge) is presented separately in Section 4.

2.2 Systems

- **TypeSafe Jev**: the System One `Choice` primitive. The state is the passage (if any) plus the question; the options are passed as the choice criteria with the instruction "Which answer is correct?"
- **Claude Haiku 4.5, Sonnet 5, Opus 5**: the Anthropic Messages API with a system prompt requesting only the answer letter, run at three reasoning settings. *None* disables extended thinking (`thinking={"type": "disabled"}`, `max_tokens=16`). *Low* and *medium* enable it via each generation's native control: gen-5 models (Sonnet 5, Opus 5) use adaptive thinking with `output_config.effort` set to `low`/`medium`, while Haiku 4.5 uses a fixed thinking budget (2,048 / 8,192 tokens). The two APIs behave differently on cost, as discussed in Section 5.

2.3 Scoring

A prediction is correct iff its parsed label equals the gold label. Claude outputs that cannot be parsed into a valid label—for example, a model that ignores the instruction and begins reasoning in the visible text until it is truncated—are counted as *unparsed* and scored *incorrect*. This is a genuine consequence of the requested "no reasoning, answer only" setup and is reported separately per run.

Cost. We meter actual token usage per request and price it at published rates (USD per 1M tokens): Jev at \$0.042 input / \$0 output; Haiku 4.5 at 1 / 5; Sonnet 5 at 2 / 10; Opus 5 at 5 / 25. Cost is reported as USD per 1,000 questions. Extended thinking bills as output tokens at the same per-model output rate, so reasoning runs cost strictly more than their no-reasoning counterparts; we price the actual thinking tokens each request emits, not the budget. Jev bills only input tokens, so its cost advantage widens with context length.

3 Results (public benchmarks)

The TypeSafe Jev row is shaded in every table; measured systems are listed in a fixed order (Haiku, Sonnet, Opus, then Jev), and within each Claude family by reasoning effort (none, low, medium).

Table 1: Accuracy by system and dataset (measured, $n = 100$, seed 42).

System	CommonsenseQA	MMLU-CF	RACE-H
Claude Haiku 4.5	88.0%	64.0%	89.0%
Claude Haiku 4.5 (low)	90.0%	74.0%	93.0%
Claude Haiku 4.5 (medium)	90.0%	75.0%	94.0%
Claude Sonnet 5	90.0%	73.0%	94.0%
Claude Sonnet 5 (low)	87.0%	76.0%	96.0%
Claude Sonnet 5 (medium)	89.0%	79.0%	97.0%
Claude Opus 5	90.0%	79.0%	95.0%
Claude Opus 5 (low)	95.0%	84.0%	96.0%
Claude Opus 5 (medium)	93.0%	82.0%	96.0%
TypeSafe Jev	91.0%	80.0%	95.0%

Table 2: Cost in USD per 1,000 questions (metered tokens × published rates).

System	CommonsenseQA	MMLU-CF	RACE-H
Claude Haiku 4.5	$0.109	$0.137	$0.506
Claude Haiku 4.5 (low)	$1.345	$1.832	$2.317
Claude Haiku 4.5 (medium)	$1.683	$2.464	$2.961
Claude Sonnet 5	$0.263	$0.339	$1.334
Claude Sonnet 5 (low)	$0.263	$0.443	$1.334
Claude Sonnet 5 (medium)	$0.263	$0.417	$1.334
Claude Opus 5	$0.658	$0.822	$3.335
Claude Opus 5 (low)	$1.123	$1.544	$3.437
Claude Opus 5 (medium)	$1.542	$2.140	$3.652
TypeSafe Jev	$0.015	$0.016	$0.031

Table 3: RACE-H: measured systems vs. published references.

System	Type	Accuracy	N	Avg latency	Cost /1k Q
Claude Haiku 4.5	measured	89.0%	100	545 ms	$0.506
Claude Haiku 4.5 (low)	measured	93.0%	100	3412 ms	$2.317
Claude Haiku 4.5 (medium)	measured	94.0%	100	4804 ms	$2.961
Claude Sonnet 5	measured	94.0%	100	1177 ms	$1.334
Claude Sonnet 5 (low)	measured	96.0%	100	1129 ms	$1.334
Claude Sonnet 5 (medium)	measured	97.0%	100	1173 ms	$1.334
Claude Opus 5	measured	95.0%	100	1662 ms	$3.335
Claude Opus 5 (low)	measured	96.0%	100	1921 ms	$3.437
Claude Opus 5 (medium)	measured	96.0%	100	1965 ms	$3.652
TypeSafe Jev	measured	95.0%	100	136 ms	$0.031
Human ceiling	published	94.5%	–	–	–
GPT-4 class	published	91.8%	–	–	–
Claude (frontier)	published	92.9%	–	–	–

Table 4: CommonsenseQA: measured systems vs. published references.

System	Type	Accuracy	N	Avg latency	Cost /1k Q
Claude Haiku 4.5	measured	88.0%	100	541 ms	$0.109
Claude Haiku 4.5 (low)	measured	90.0%	100	2730 ms	$1.345
Claude Haiku 4.5 (medium)	measured	90.0%	100	3455 ms	$1.683
Claude Sonnet 5	measured	90.0%	100	1185 ms	$0.263
Claude Sonnet 5 (low)	measured	87.0%	100	1191 ms	$0.263
Claude Sonnet 5 (medium)	measured	89.0%	100	1223 ms	$0.263
Claude Opus 5	measured	90.0%	100	1677 ms	$0.658
Claude Opus 5 (low)	measured	95.0%	100	1825 ms	$1.123
Claude Opus 5 (medium)	measured	93.0%	100	2016 ms	$1.542
TypeSafe Jev	measured	91.0%	100	149 ms	$0.015
Human	published	88.9%	–	–	–
GPT-4 class	published	79.0%	–	–	–
Claude (frontier)	published	78.0%	–	–	–

Table 5: MMLU-CF: measured systems vs. published references.

System	Type	Accuracy	N	Avg latency	Cost /1k Q
Claude Haiku 4.5	measured	64.0%	100	569 ms	$0.137
Claude Haiku 4.5 (low)	measured	74.0%	100	3374 ms	$1.832
Claude Haiku 4.5 (medium)	measured	75.0%	100	4481 ms	$2.464
Claude Sonnet 5	measured	73.0%	100	1238 ms	$0.339
Claude Sonnet 5 (low)	measured	76.0%	100	1290 ms	$0.443
Claude Sonnet 5 (medium)	measured	79.0%	100	1284 ms	$0.417
Claude Opus 5	measured	79.0%	100	1609 ms	$0.822
Claude Opus 5 (low)	measured	84.0%	100	1981 ms	$1.544
Claude Opus 5 (medium)	measured	82.0%	100	2168 ms	$2.140
TypeSafe Jev	measured	80.0%	100	140 ms	$0.016
Human (est.)	published	90.0%	–	–	–
GPT-4 class	published	73.4%	–	–	–
Claude (frontier)	published	70.0%	–	–	–

4 Grace-Coolidge: a contamination-controlled corpus

To control for the possibility that public test items sit in pretraining data, we hand-authored a 40-question corpus (20 easy, 20 hard) over the full Wikipedia article on Grace Coolidge (∼20k tokens), each question with six options. The whole article is supplied as context on every call and every distractor is an in-passage, semantically valid name or value, so answering requires reading the passage rather than recalling a memorized item. The hard split adds negation, temporal-ordering, and inference questions, several of them deliberately unanswerable (the correct answer is "Not enough information").

Table 6: Grace-Coolidge ($n = 40$): measured systems. Cost reflects the ∼20k-token article supplied on every call.

System	Type	Accuracy	N	Avg latency	Cost /1k Q
Claude Haiku 4.5	measured	70.0%	40	962 ms	$18.523
Claude Haiku 4.5 (low)	measured	90.0%	40	3007 ms	$19.787
Claude Haiku 4.5 (medium)	measured	90.0%	40	3431 ms	$20.003
Claude Sonnet 5	measured	87.5%	40	1548 ms	$51.879
Claude Sonnet 5 (low)	measured	87.5%	40	1524 ms	$51.912
Claude Sonnet 5 (medium)	measured	85.0%	40	1472 ms	$51.899
Claude Opus 5	measured	92.5%	40	2252 ms	$129.690
Claude Opus 5 (low)	measured	90.0%	40	1960 ms	$129.952
Claude Opus 5 (medium)	measured	90.0%	40	2181 ms	$130.213
TypeSafe Jev	measured	90.0%	40	429 ms	$0.840

Table 7: Grace-Coolidge accuracy split by difficulty.

System	Easy	Hard	Overall	Cost /1k Q
Claude Haiku 4.5	70.0%	70.0%	70.0%	$18.523
Claude Haiku 4.5 (low)	100.0%	80.0%	90.0%	$19.787
Claude Haiku 4.5 (medium)	100.0%	80.0%	90.0%	$20.003
Claude Sonnet 5	100.0%	75.0%	87.5%	$51.879
Claude Sonnet 5 (low)	100.0%	75.0%	87.5%	$51.912
Claude Sonnet 5 (medium)	100.0%	70.0%	85.0%	$51.899
Claude Opus 5	100.0%	85.0%	92.5%	$129.690
Claude Opus 5 (low)	100.0%	80.0%	90.0%	$129.952
Claude Opus 5 (medium)	100.0%	80.0%	90.0%	$130.213
TypeSafe Jev	100.0%	80.0%	90.0%	$0.840

On the easy split, every system except no-reasoning Haiku 4.5 scores 100%; the hard split is the discriminator. Jev (90.0% overall) beats every Sonnet 5 and matches every reasoning-enabled Opus 5 configuration, trailing only no-reasoning Opus 5 by a single hard question—at roughly 1/150 of Opus 5's cost. Reasoning does not help the frontier models here: Sonnet 5 and Opus 5 stay flat or dip slightly, because the long article already forces them to read rather than recall. The one place reasoning matters is Haiku 4.5, whose no-reasoning 70% (7 truncated, unparsed answers) recovers to 90% once thinking is enabled. Crucially, because the article dominates the token bill, turning on reasoning barely moves cost (Opus 5 $129.69 → $130.21), so the frontier models pay two orders of magnitude more than Jev regardless of reasoning.

5 Analysis

Accuracy (no reasoning). Among the no-reasoning, answer-only configurations, TypeSafe Jev is the top measured system on CommonsenseQA (91.0%) and on the contamination-free MMLU-CF (80.0%), ties Opus 5 on RACE-H (95.0%, at the human ceiling), and trails only Opus 5 on Grace-Coolidge (90.0% vs. 92.5%, one hard question apart). It clears the GPT-4-class published

number on all three public benchmarks.

Reasoning effort. Turning on reasoning helps the frontier models most on the quantitative MMLU-CF split—Opus 5 rises from 79.0% to 84.0% (low), Sonnet 5 from 73.0% to 79.0% (medium), Haiku 4.5 from 64.0% to 75.0% (medium)—and lets reasoning-enabled Opus 5 edge past Jev on CommonsenseQA (95.0% vs. 91.0%) and MMLU-CF (84.0% vs. 80.0%), and Sonnet 5 past Jev on RACE-H (97.0% vs. 95.0%). The gains are one to five points and come at a steep multiple: on the public benchmarks Opus 5's cost roughly doubles (e.g. MMLU-CF \$0.82 → \$2.14) and its latency stays ∼2 s, while Jev delivers within two points at 40–100× less cost and ∼10× less latency. The two thinking APIs also price very differently: gen-5 *adaptive* thinking (Sonnet 5, Opus 5) spends only the tokens a question needs—often near zero on easy items, so Sonnet 5's cost is essentially unchanged—whereas Haiku 4.5's fixed-*budget* thinking emits tokens up to the budget even on trivial questions, inflating its cost more than tenfold (CommonsenseQA \$0.11 → \$1.35) for a two-point gain.

Latency. Across datasets Jev answers in roughly 136–429 ms, versus ∼550 ms for no-reasoning Haiku 4.5, ∼1.2 s for Sonnet 5, and ∼1.6–2.3 s for Opus 5; enabling reasoning adds little for the gen-5 models (adaptive thinking is near-instant on these items) but roughly doubles Haiku's latency. Jev keeps an order-of-magnitude speed advantage at comparable accuracy.

Cost. The gap is widest exactly where it matters for production classification: long context. On Grace-Coolidge, where every call carries a ∼20k-token article, Jev costs \$0.84 per 1,000 questions versus \$129.69 for Opus 5 and \$51.88 for Sonnet 5—roughly 150× and 60× cheaper respectively—because Jev bills only input tokens at \$0.042/1M and charges nothing for output, while chat models add \$5–\$25/1M output on top of costlier input. Even on the short public benchmarks Jev is 40–100× cheaper than Opus 5.

The no-reasoning constraint. The constraint is close to free on passage-grounded RACE-H but costly on quantitative MMLU-CF questions. With thinking disabled and output capped at a letter, Haiku 4.5 (7 items) and Sonnet 5 (3 items) began working problems in the visible text and were truncated, yielding unparsed—and therefore incorrect—answers; the same effect drives Haiku's 70% on Grace-Coolidge (7 unparsed). Opus 5 answered directly and did not degrade. Enabling reasoning removes the penalty—the models emit thinking tokens instead of leaking work into the answer channel, which is why Haiku recovers to 90% on Grace-Coolidge—but this is a workaround that reintroduces the very cost and latency the constraint was meant to bound. Because the `Choice` primitive returns a label by construction, TypeSafe never pays this penalty in either setting.

Contamination. The relative ordering is stable on the two contamination controls—MMLU-CF (contamination-free) and Grace-Coolidge (a single long article no model could have "memorized the test" for). That the pattern holds on both is evidence the comparison is not an artifact of memorized public test items.

6 Conclusion

Under a single fixed, reproducible protocol, TypeSafe Jev is competitive with the frontier Claude generation on multiple-choice accuracy while being approximately 10× faster and, on long-context items, over 100× cheaper. It leads every no-reasoning chat configuration on two of three public

benchmarks; enabling low/medium reasoning buys the frontier models a one-to-five point edge on the public splits, but at two-to-twelve times the cost and $\sim$10$\times$ the latency, and with no benefit on the long-context Grace-Coolidge corpus. Jev is also unaffected by the no-reasoning / answer-only constraint that degrades smaller and mid-tier chat models on quantitative items. Across all reasoning settings it remains the best accuracy-per-dollar and accuracy-per-latency system measured. All results are regenerated by the scripts in this repository; see the top-level `README.md` to reproduce.

7 Appendix: example prompts

One worked example per dataset, generated from the first preprocessed item by `build_appendix.py`. Each shows the TypeSafe request payload and the Claude text prompt for the same question.

RACE-H

Gold answer: `B`.

TypeSafe payload.

```
{
  "state": "Passage:\nIn the dark and damp market,Chen ShuChu,near her sixties,owns a stall that
    her father left her.The stall,called YuanJin Vegetables,is her everything.Chen earns only a
    little money by selling vegetables,but she has donated about $321,550 to help poor children.\
    nIn March,Forbes magazine named her one of the 48 great philanthropists from the AsiaPacific
    region.A month later,TIME magazine selected the year's top 100 influential people and Chen
    was one of them.Although she has received many honours and rewards,Chen only cares about her
    vegetable stall and whether her regular customers buy her vegeta\n... [passage truncated for
    print; 1,458 chars total] ...\n\nQuestion: Why has Chen been called \"market manager\"?",
  "model": "jev-latest",
  "questions": {
    "answer": {
      "type": "choice",
      "instructions": "Which answer is correct?",
      "criteria": {
        "A": "Because she has received many honours and rewards.",
        "B": "Because she often arrives early and leaves late in the market.",
        "C": "Because she has donated much money to the market.",
        "D": "Because her vegetable stall is the biggest one in the market."
      }
    }
  }
}
```

Claude prompt.

```
[system]
You are answering a multiple-choice question. Respond with ONLY the single letter of the correct
    option (one of the provided labels) and nothing else. No explanation.

[user]
Passage:
In the dark and damp market,Chen ShuChu,near her sixties,owns a stall that her father left her.
    The stall,called YuanJin Vegetables,is her everything.Chen earns only a little money by
    selling vegetables,but she has donated about $321,550 to help poor children.
In March,Forbes magazine named her one of the 48 great philanthropists from the AsiaPacific
    region.A month later,TIME magazine selected the year's top 100 influential people and Chen
    was one of them.Although she has received many honours and rewards,Chen only cares about her
    vegetable stall and whether her regular customers buy her vegeta
```

```
... [passage truncated for print; 1,458 chars total] ...

Question: Why has Chen been called "market manager"?

Options:
A. Because she has received many honours and rewards.
B. Because she often arrives early and leaves late in the market.
C. Because she has donated much money to the market.
D. Because her vegetable stall is the biggest one in the market.
```

CommonsenseQA

Gold answer: `A`.

TypeSafe payload.

```
{
  "state": "Question: You can do knitting to get the feeling of what?",
  "model": "jev-latest",
  "questions": {
    "answer": {
      "type": "choice",
      "instructions": "Which answer is correct?",
      "criteria": {
        "A": "relaxation",
        "B": "arthritis",
        "C": "adrenaline",
        "D": "your",
        "E": "sweater may produced"
      }
    }
  }
}
```

Claude prompt.

```
[system]
You are answering a multiple-choice question. Respond with ONLY the single letter of the correct
    option (one of the provided labels) and nothing else. No explanation.

[user]
Question: You can do knitting to get the feeling of what?

Options:
A. relaxation
B. arthritis
C. adrenaline
D. your
E. sweater may produced
```

MMLU-CF

Gold answer: `C`.

TypeSafe payload.

```
{
  "state": "Question: Where are the income and expenses related to the prize fund displayed?",
  "model": "jev-latest",
```

```
  "questions": {
    "answer": {
      "type": "choice",
      "instructions": "Which answer is correct?",
      "criteria": {
        "A": "None of the other choices",
        "B": "Cash Account",
        "C": "Income and Expenditure Account",
        "D": "Assets side of the Balance Sheet"
      }
    }
  }
}
```

Claude prompt.

```
[system]
You are answering a multiple-choice question. Respond with ONLY the single letter of the correct
    option (one of the provided labels) and nothing else. No explanation.

[user]
Question: Where are the income and expenses related to the prize fund displayed?

Options:
A. None of the other choices
B. Cash Account
C. Income and Expenditure Account
D. Assets side of the Balance Sheet
```

Grace-Coolidge

Gold answer: `C`.

TypeSafe payload.

```
{
  "state": "Passage:\n'''Grace Anna Coolidge''' ([[Maiden and married names|ne]] '''Goodhue''';
    January 3, 1879 - July 8, 1957) was [[first lady of the United States]] from 1923 to 1929 as
    the wife of the 30th [[president of the United States]], [[Calvin Coolidge]]. She was
    previously the [[second lady of the United States]] from 1921 to 1923 and the [[List of First
     Spouses of Massachusetts|first lady of Massachusetts]] from 1919 to 1921.\n\nCoolidge was
    born and raised in [[Burlington, Vermont]], and attended the [[University of Vermont]] where
    she co-founded the school's chapter of [[Pi Beta Phi]]. She moved to [[\n... [passage
    truncated for print; 63,022 chars total] ...\n\nQuestion: Who funded the portrait that was
    later moved to the Red Room by Jacqueline Kennedy?",
  "model": "jev-latest",
  "questions": {
    "answer": {
      "type": "choice",
      "instructions": "Which answer is correct?",
      "criteria": {
        "A": "The Republican Party",
        "B": "Howard Chandler Christy",
        "C": "Pi Beta Phi",
        "D": "Frank Stearns",
        "E": "The Coolidge Foundation",
        "F": "None of the above"
      }
    }
```

```
  }
}
```

Claude prompt.

```
[system]
You are answering a multiple-choice question. Respond with ONLY the single letter of the correct
    option (one of the provided labels) and nothing else. No explanation.

[user]
Passage:
'''Grace Anna Coolidge''' ([[Maiden and married names|ne]] '''Goodhue'''; January 3, 1879 - July
    8, 1957) was [[first lady of the United States]] from 1923 to 1929 as the wife of the 30th [[
    president of the United States]], [[Calvin Coolidge]]. She was previously the [[second lady
    of the United States]] from 1921 to 1923 and the [[List of First Spouses of Massachusetts|
    first lady of Massachusetts]] from 1919 to 1921.

Coolidge was born and raised in [[Burlington, Vermont]], and attended the [[University of Vermont
    ]] where she co-founded the school's chapter of [[Pi Beta Phi]]. She moved to [[
... [passage truncated for print; 63,022 chars total] ...

Question: Who funded the portrait that was later moved to the Red Room by Jacqueline Kennedy?

Options:
A. The Republican Party
B. Howard Chandler Christy
C. Pi Beta Phi
D. Frank Stearns
E. The Coolidge Foundation
F. None of the above
```