

Cairo Rosetta versus Frontier Language Models in Predicting Individual Human Choices

First discrete-choice experiment in Rosetta Arena

Cairo Labs · Cairo AI

October 2026 · Evaluated run: October 4, 2026, 23:44

Abstract. We present the first experiment of Rosetta Arena, a benchmarking environment designed to measure how well different artificial intelligence systems predict the exact decision a real person made when facing a closed set of alternatives. We compare Cairo Rosetta, a specialized behavior model developed by Cairo AI, against three general-purpose language models accessed through OpenRouter (Claude Sonnet 5.5, GPT-5.6 Terra, and Gemini 3.8 Flash) and against a frequentist baseline (historical per-event mode), on 30 real decisions drawn from the validation and test partitions of the Twin-2K-500 and CaSiNo datasets. Rosetta achieved the highest point accuracy (60.0%; 95% confidence interval [CI]: 42.3–75.4%), followed by Claude Sonnet 5.5 (56.7%), GPT-5.6 Terra (50.0%), the baseline (40.0%), and Gemini 3.8 Flash (36.7%). At the current sample size, the accuracy advantage over the best language model is not statistically distinguishable. The clearest difference is in efficiency: Rosetta reduced mean latency by 41% relative to the fastest language models (1.73 s vs. 2.93–2.94 s) and by 63% relative to Gemini 3.8 Flash, without incurring API charges, compared with USD 1.90–3.03 per 1,000 predictions for the commercial models. In point estimates, Rosetta Pareto-dominates all three language models evaluated. The results are preliminary and warrant a larger-scale evaluation with paired, case-level analysis.

Keywords: behavioral prediction · discrete choice · *Large Behavior Models* · language models · benchmarking · inference efficiency

1 Introduction

Anticipating which decision a specific person will make—not just how a population behaves on average—is a central problem in organizational psychology, behavioral economics, and product design. In recent years, large language models (LLMs) have been used as simulators of human subjects, with promising results at the aggregate (Argyle et al., 2023; Horton, 2023) and individual levels (Park et al., 2024), although individual-level validation has been limited by the scarcity of high-quality public behavioral data (Toubia et al., 2025). In parallel, models trained specifically to predict human behavior and cognition have emerged (Binz et al., 2025), raising a practical question: can a specialized model compete with frontier LLMs on accuracy while outperforming them on cost and speed?

Cairo Rosetta (hereafter, Rosetta) is a specialized behavior model—a *Large Behavior Model* (LBM)—developed by Cairo AI. Its goal is to estimate, from a person's psychometric profile and the characteristics of a situation, which action that person is most likely to take. Unlike general-purpose LLMs, Rosetta is designed to produce structured predictions with low latency and low marginal cost, in order to enable large-scale behavioral simulation.

This report documents the first experiment of Rosetta Arena, our benchmarking environment. Its contributions are threefold: (i) a blind, leak-free evaluation protocol for predicting individual discrete choices; (ii) a direct comparison of Rosetta with three frontier LLMs and a frequentist baseline on accuracy, latency, and cost; and (iii) an explicit analysis of the statistical uncertainty associated with sample size, which delimits which conclusions can—and cannot—be drawn from this first run.

2 Evaluated Systems

Five competitors were evaluated, as summarized in Table 1.

Table 1. Evaluated competitors.

Competitor	Type	Access	Identifier
Cairo Rosetta	Specialized behavior model (LBM)	Dedicated inference	<code>cairo_rosetta</code>
Claude Sonnet 5.5	General-purpose LLM	API via OpenRouter	Claude Sonnet 5.5 (Anthropic)
GPT-5.6 Terra	General-purpose LLM	API via OpenRouter	<code>openai/gpt-5.6-terra</code>
Gemini 3.8 Flash	General-purpose LLM	API via OpenRouter	<code>google/gemini-3.8-flash</code>
Per-event average	Frequentist baseline (historical mode)	Local computation	Deterministic lookup

Cairo Rosetta. A proprietary Cairo AI model specialized in predicting individual behavior. It receives the structured description of the person and the situation and directly returns the catalog alternative that best matches the profile. Because it is a proprietary development, details of its architecture, internal representation, and training procedure are deliberately omitted: in this report, Rosetta is evaluated as a black box, under the same input conditions as the other competitors.

Frontier language models. Three commercial general-purpose LLMs accessed through OpenRouter. Each receives a structured prompt with the person's profile, the description of the situation, and the numbered menu of options, and must return, in JSON format, the option it predicts the subject chose.

Per-event average. A deterministic frequentist baseline that predicts, for each event, the option chosen most frequently in the reference data. Despite its name in the system log, it does not compute an average but the event's historical mode. In the absence of individual information, predicting the mode maximizes expected accuracy; this baseline therefore serves as a minimum threshold: a system that does not beat it is not extracting useful information from the person's profile.

3 Experimental Design

3.1 Data sources and sampling

Cases come from empirical decisions made by real people in economic, negotiation, and social dilemmas, drawn from two public datasets: Twin-2K-500 (Toubia et al., 2025), which gathers the answers of more than 2,000 participants to more than 500 questions, including psychometric measures and behavioral tasks, and CaSiNo (Chawla et al., 2021), a corpus of negotiation dialogues about splitting resources for a camping trip.

To ensure a leak-free evaluation, cases were drawn exclusively from out-of-training partitions (`val` and `test`) or from held-out control events (*holdout*); no evaluated case was part of Rosetta's training data or of the baseline's reference data. Only situations with a closed catalog of 2 to 12 discrete alternatives (`option_key`) were included. The order of alternatives was randomized with a fixed pseudorandom seed (`seed`) to neutralize position biases, such as a model's systematic preference for the first or last option. The final sample consists of $N = 30$ decisions.

3.2 Case structure

Each test unit provides all competitors with the same information, organized into three components. The *individual profile* contains the person's standardized cognitive and personality traits, expressed as normalized scores (z) or as equivalent semantic calibrations. The *situation context* includes a textual description of the scenario together with objective contextual descriptors (for example, what is at stake, the degree of uncertainty, and the conditions of the interaction). The *option catalog* lists the discrete actions available to the person at that moment. The participant's actual decision (`real_key`) remains hidden during inference and is used only at the scoring stage.

3.3 Inference protocol

Competitors were run independently and in parallel on each case, against the same catalog and in the same randomized order. The LLMs received the case through the structured prompt described in Section 2; the baseline deterministically selected the event's historical mode; and Rosetta received the case and directly resolved which catalog alternative matches the person's profile. In every case, the output is a single option from the catalog, which allows exact and uniform scoring.

3.4 Metrics

The primary metric is accuracy, defined from the exact hit on each case i , where $\hat{y}_i$ is the predicted option and y_i the option the person chose:

$$\text{Hit}_i = 1 \text{ if } \hat{y}_i = y_i, \text{ and } \text{Hit}_i = 0 \text{ otherwise} \quad (1)$$

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \text{Hit}_i \quad (2)$$

As a complementary metric, we report the mean absolute error (MAE). In a discrete-choice problem with binary scoring, the MAE equals the error rate and therefore carries no information independent of accuracy; it is retained for consistency with the Rosetta Arena log:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N (1 - \text{Hit}_i) = 1 - \text{Accuracy} \quad (3)$$

For each prediction we also recorded end-to-end latency (`latencia_ms`), measured from query dispatch to receipt of the predicted option, and the direct financial cost of inference (`costo`), corresponding to external API charges. Efficiency comparisons are expressed as relative latency reduction and speedup, where L_R and L_k are the mean latencies of Rosetta and of competitor k :

$$\Delta_{\text{lat}} = 1 - \frac{L_R}{L_k}, \quad \text{speedup} = \frac{L_k}{L_R} \quad (4)$$

3.5 Statistical analysis

Uncertainty in each accuracy estimate is quantified with 95% Wilson confidence intervals (Wilson, 1927), which are more appropriate than the normal approximation for small samples. Because all competitors were evaluated on the same 30 cases, the appropriate pairwise test is the exact McNemar test (McNemar, 1947), which requires the case-by-case concordance table. Since this report relies on aggregate results, we present (i) a two-proportion z test, which ignores pairing, and (ii) the range of McNemar p -values compatible with the observed totals, obtained by enumerating all feasible concordance tables. This range bounds the significance that a case-level analysis could achieve.

3.6 Execution infrastructure

The experiment ran on the Rosetta Arena backend as an asynchronous process of chained atomic batches (*self-chaining ticks*). Each run is logged in `rosetta_arena_corridas` and each individual attempt in `rosetta_arena_intentos`, with each competitor's exact prediction, latency, and cost. The system includes fault tolerance and exponential-backoff retries on network interruptions, which, together with the fixed randomization seed, makes the run reproducible. The analyzed experiment was run on October 4, 2026, at 23:44 and validated this infrastructure end to end, allowing the evaluation to be repeated and expanded continuously as Rosetta evolves.

4 Results

4.1 Accuracy

Table 2 reproduces the results recorded in the database. Rosetta achieved the highest accuracy (18/30; 60.0%) and the lowest error (0.40), followed by Claude Sonnet 5.5 (17/30; 56.7%) and GPT-5.6 Terra (15/30; 50.0%). The baseline got 12 of 30 cases right (40.0%), and Gemini 3.8 Flash had the lowest performance (11/30; 36.7%), below the baseline.

Table 2. Results of the discrete-choice run ($N = 30$), sorted by accuracy.

Competitor	Hits	Accuracy	95% CI (Wilson)	Error (MAE)	Mean latency	Total cost
Cairo Rosetta	18 / 30	60.0%	42.3–75.4%	0.40	1,733 ms	USD 0.0000
Claude Sonnet 5.5	17 / 30	56.7%	39.2–72.6%	0.43	2,937 ms	USD 0.0908
GPT-5.6 Terra	15 / 30	50.0%	33.2–66.8%	0.50	2,930 ms	USD 0.0571
Per-event average	12 / 30	40.0%	24.6–57.7%	0.60	0 ms	USD 0.0000
Gemini 3.8 Flash	11 / 30	36.7%	21.9–54.5%	0.63	4,728 ms	USD 0.0635

Note: MAE equals $1 - \text{accuracy}$ in discrete choice (equation 3). Total cost corresponds to external API charges; the baseline recorded 0 ms latency because it is a precomputed lookup.

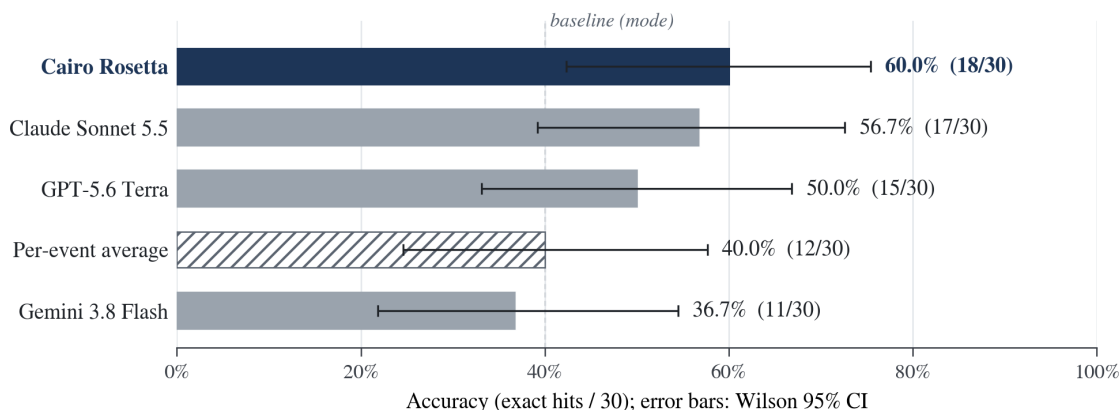


Figure 1. Accuracy by competitor with 95% Wilson confidence intervals. The dashed line marks the baseline accuracy (historical mode). The width of the intervals ($\approx \pm 16.5$ pp) reflects the sample size of this first run.

Confidence intervals are wide—about ± 16.5 percentage points (pp)—and overlap substantially (Figure 1). Table 3 summarizes Rosetta's comparison against each competitor. The advantage over Claude Sonnet 5.5 amounts to a single case and is not significant under any feasible configuration ($p = 1.00$); against GPT-5.6 Terra, the 10 pp difference (3 cases) is not significant either (McNemar p between 0.25 and 0.70). The larger differences—20 pp over the baseline and 23.3 pp over Gemini 3.8 Flash—do not reach significance in the unpaired test ($p = 0.12$ and 0.07), but could do so in a paired analysis, whose feasible range is 0.031–0.36 and 0.016–0.26, respectively. The lower end corresponds to the scenario in which Rosetta's hits include all of the competitor's hits; determining where the actual result falls requires the case-level concordance table.

4.2 Hits on non-modal decisions

The baseline enables an additional reading of the results. By construction, its 12 hits correspond exactly to the cases in which the person chose the historically most frequent option; in the remaining 18 cases, the individual departed from the mode. Consequently, any competitor with k total hits must have at least $k - 12$ hits on these non-modal decisions. Rosetta correctly anticipated at least 6 of the 18 atypical decisions ($\geq 33\%$), a type of hit the baseline cannot achieve; the corresponding bounds are 5 for Claude Sonnet 5.5, 3 for GPT-5.6 Terra, and 0 for Gemini 3.8 Flash.

This bound is consistent with the hypothesis that Rosetta extracts useful information from the individual profile, although the exact magnitude can only be established through case-level analysis.

Table 3. Comparison of Rosetta against each competitor. Δ : accuracy difference in Rosetta’s favor. The z test ignores pairing; the McNemar range covers all concordance tables compatible with the observed hits. The last column is a lower bound derived from the totals (Section 4.2).

Competitor	Accuracy	Δ vs. Rosetta	Δ cases	z -test p (unpaired)	McNemar p (feasible range)	Min. hits on non-modal cases
Cairo Rosetta	60.0%	—	—	—	—	≥ 6 of 18
Claude Sonnet 5.5	56.7%	+3.3 pp	+1	0.79	1.00	≥ 5 of 18
GPT-5.6 Terra	50.0%	+10.0 pp	+3	0.44	0.25–0.70	≥ 3 of 18
Per-event average	40.0%	+20.0 pp	+6	0.12	0.031–0.36	0 of 18 (by construction)
Gemini 3.8 Flash	36.7%	+23.3 pp	+7	0.07	0.016–0.26	≥ 0 of 18

4.3 Latency

Rosetta recorded a mean latency of 1,733 ms, versus 2,930 ms for GPT-5.6 Terra, 2,937 ms for Claude Sonnet 5.5, and 4,728 ms for Gemini 3.8 Flash (Figure 2a; Table 4). This represents a 40.9–41.0% reduction relative to the two fastest LLMs—a 1.69 $\times$ speedup—and a 63.3% reduction relative to Gemini 3.8 Flash (2.73 $\times$). The baseline, being a lookup in a precomputed table, runs with practically zero latency and is excluded from this comparison.

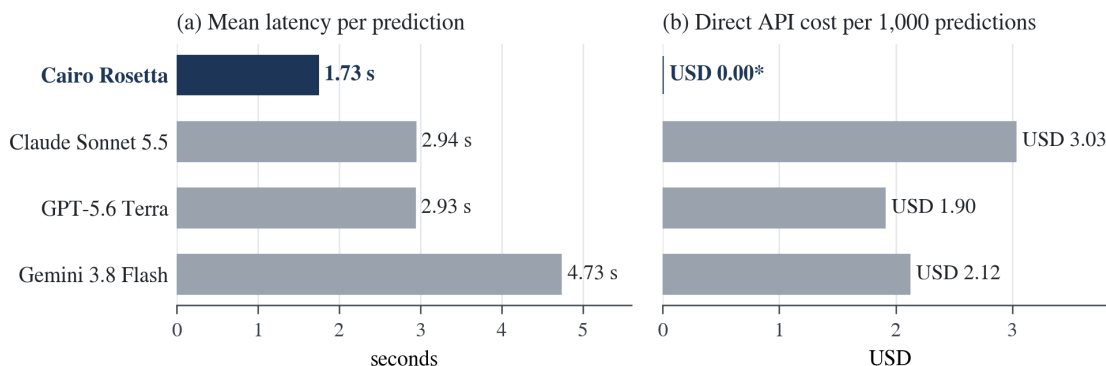


Figure 2. Inference efficiency. (a) Mean end-to-end latency per prediction. (b) Direct API cost normalized to 1,000 predictions. *Rosetta incurred no external API charges; the cost of its dedicated infrastructure is not captured in this telemetry (see Section 6).

4.4 Cost

Rosetta incurred no external API charges during the experiment, while the LLMs accumulated between USD 0.057 and USD 0.091 over 30 predictions. Normalized, this equals USD 3.03 (Claude Sonnet 5.5), USD 2.12 (Gemini 3.8 Flash), and USD 1.90 (GPT-5.6 Terra) per 1,000 predictions (Figure 2b). Although the absolute figures are small at this scale, they grow linearly with volume: assuming similar prompt sizes, a simulation of one million predictions—for example, 10,000 profiles exposed to 100 scenarios—would cost roughly USD 1,900 to USD 3,030 on commercial APIs, versus zero API cost for Rosetta. Measured per correct prediction, the cost of the LLMs ranged from USD 0.0038 (GPT-5.6 Terra) to USD 0.0058 (Gemini 3.8 Flash), as detailed in Table 4.

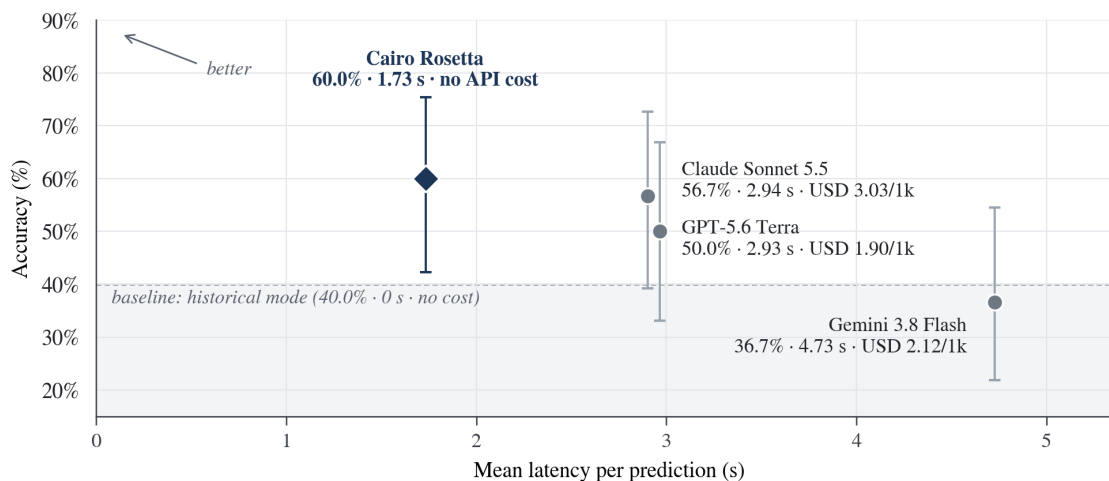
Table 4. Derived efficiency metrics. Reductions and speedups computed relative to Rosetta using equation (4); the one-million projection assumes a constant per-call cost.

Competitor	Latency reduction (Rosetta)	Speedup	Cost per prediction	Cost per 1,000 predictions	Projection to 10 ⁶ predictions	Cost per hit
Cairo Rosetta	reference	1.00×	USD 0*	USD 0*	USD 0*	USD 0*
Claude Sonnet 5.5	41.0%	1.69×	USD 0.0030	USD 3.03	USD 3,027	USD 0.0053
GPT-5.6 Terra	40.9%	1.69×	USD 0.0019	USD 1.90	USD 1,903	USD 0.0038
Gemini 3.8 Flash	63.3%	2.73×	USD 0.0021	USD 2.12	USD 2,117	USD 0.0058

*No external API charges. Excludes compute, hosting, and amortized training of Rosetta's dedicated infrastructure.

4.5 Synthesis: Pareto dominance

Figure 3 places each competitor on the accuracy–latency plane. In point estimates, Rosetta Pareto-dominates the three LLMs evaluated: it simultaneously offers higher accuracy, lower latency, and lower direct cost. The only non-dominated alternative is the baseline, which is faster and free but 20 pp less accurate. The asymmetry of the evidence should be emphasized: dominance in accuracy is subject to the uncertainty described in Section 4.1, whereas the latency and cost differences are substantial in magnitude.

**Figure 3.** Accuracy–latency plane with 95% Wilson intervals; the upper-left corner is preferable. The shaded band indicates performance below the baseline. The points for Claude Sonnet 5.5 and GPT-5.6 Terra, with nearly identical latencies, are offset horizontally by ± 0.035 s for readability.

5 Discussion and First Impressions

5.1 Competitive parity with frontier models

Our first impression is that a specialized model can match—and, in this run, marginally exceed—frontier language models on an admittedly difficult task: predicting the choice of a specific person. The result is relevant because the LLMs evaluated are large-scale general-purpose systems, while Rosetta is designed specifically for this class of problem. Nonetheless, with 30 cases the result should be read as competitive parity with a directional advantage in Rosetta's favor, not as demonstrated superiority.

5.2 Efficiency as a structural advantage

The experiment's strongest advantage lies not in accuracy but in efficiency. Whereas the accuracy differences hinge on a few cases, the latency differences (about 1.2 s per prediction versus the fastest LLMs) and cost differences (no API charges versus USD 1.90–3.03 per 1,000 predictions) are large and stem from structural differences between a specialized model with dedicated inference and a general-purpose LLM served through an API. For the use case that motivates Rosetta—simulating the responses of thousands or millions of profiles across multiple scenarios—these differences translate directly into economic viability and the ability to deliver results at interactive speed.

5.3 The baseline as a test of individualization

That the historical mode is correct in only 40% of cases indicates that the sample contains a considerable share of decisions that depart from what the majority does; the task therefore rewards the use of individual information. In this context, it is notable that Gemini 3.8 Flash did not beat the baseline: being a frontier model does not, by itself, guarantee extracting useful signal from a person's profile. The spread among LLMs (36.7% to 56.7%) further suggests that generalist performance on this task is sensitive to model choice, although this spread is also affected by sample size.

6 Limitations and Threats to Validity

Sample size and case-level analysis. With $N = 30$, each accuracy estimate has a margin of roughly ± 16.5 pp. Under an unpaired approximation, detecting a 10 pp difference (60% vs. 50%) with 80% power would require about 390 cases per competitor, and a 3.3 pp difference like the one observed against Claude Sonnet 5.5, about 3,500. A paired design reduces these requirements but needs the case-level concordance table, available in `rosetta_arena_intentos` and not included in this report.

Domain advantage. Rosetta and the baseline have training or reference data from the same sources as the evaluated cases (in different partitions), whereas the LLMs operate without adaptation (*zero-shot*). The comparison is therefore that of a domain specialist versus generalists; Rosetta's generalization to unseen data sources remains to be evaluated.

Possible LLM contamination. Twin-2K-500 and CaSiNo are public, so it cannot be ruled out that they are part of the commercial LLMs' pretraining corpora. If present, this contamination would favor the LLMs, not Rosetta.

Prompting protocol. The LLMs were evaluated with a single structured prompt. Variants such as in-context examples (*few-shot*), extended reasoning modes, or sampling-parameter adjustments could change their accuracy as well as their latency and cost.

Comparability of efficiency. LLM latency includes OpenRouter routing and each provider's network, while Rosetta runs on dedicated infrastructure; part of the difference is attributable to infrastructure and not only to the model. The cost recorded for Rosetta excludes compute, hosting, and amortized training, so a total-cost-of-ownership comparison requires instrumenting those components. In addition, latencies are reported as means, without their distribution (p50/p95).

Sample composition and single run. The number of alternatives varies from 2 to 12 per case, so the chance level differs across cases, and the share of cases from each source is not broken down here. Moreover, the results correspond to a single run with a single seed; variability across seeds and the inherent non-determinism of the LLMs have not been quantified.

7 Future Work

The next iteration of Rosetta Arena should: (i) expand the sample to several hundred cases, stratified by source and dilemma type (economic, negotiation, and social); (ii) report the case-level concordance table, exact McNemar tests, and accuracy separately for modal and non-modal decisions; (iii) add probabilistic metrics (log-loss, Brier score, and calibration) and chance-adjusted accuracy; (iv) evaluate the LLMs with more demanding prompting protocols and include other specialized models as competitors; (v) measure Rosetta's generalization on data sources not seen during training, including domains closer to its intended use cases, such as reactions to messages and communications; and

(vi) report latency distributions and the total cost of inference, including in-house infrastructure. An ablation experiment in which all competitors receive the cases without the individual profile would also quantify how much of the accuracy comes from psychometric information.

8 Conclusions

This first Rosetta Arena experiment supports four preliminary conclusions:

1. **Competitive accuracy.** Rosetta performs on par with the frontier LLMs evaluated and achieved the best point result (60.0%), with advantages of 3.3 pp over Claude Sonnet 5.5 and 10 pp over GPT-5.6 Terra. With $N = 30$, these differences are not statistically conclusive.
2. **Individual signal.** Rosetta beat the baseline by 20 pp and correctly anticipated at least one third of the decisions that departed from the majority, preliminary evidence that it leverages profile information. Gemini 3.8 Flash did not beat that baseline.
3. **Superior efficiency.** The clearest advantage is operational: Rosetta responded in 1.73 s—41% less than the fastest LLMs and 63% less than Gemini 3.8 Flash—with no API charges, versus USD 1.90–3.03 per 1,000 predictions. At the scale of millions of predictions, this difference is decisive for the viability of massive behavioral simulations.
4. **Pareto dominance.** In point estimates, Rosetta dominates the three LLMs evaluated on accuracy, latency, and direct cost. The results warrant a larger-scale evaluation, with paired case-level analysis, before making definitive claims of superior accuracy.

In short, this first run shows competitive parity in accuracy, with a directional advantage for Rosetta, and a clear and substantial advantage in speed and cost.

References

- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.
- Binz, M., Akata, E., Bethge, M., et al. (2025). A foundation model to predict and capture human cognition. *Nature*.
- Chawla, K., Ramirez, J., Clever, R., Lucas, G., May, J., & Gratch, J. (2021). CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of NAACL-HLT 2021*.
- Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper* No. 31122.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv:2411.10109*.
- Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*. <https://doi.org/10.1287/mksc.2025.0262> (preprint: arXiv:2505.17479).
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.