Evaluating a College Admissions Chancing Prompt

Corrected results and a frozen fresh-set comparison

Aristo Scholars (aristoscholars.com). October 2026. Research edition; no live product deployment.

Abstract

We evaluated a school-informed language-model prompt against self-reported college admissions outcomes. The original Reddit corpus contains 1,124 profiles, split into 700 development profiles and 424 test profiles. Of the test profiles, 423 contribute 6,129 scorable decisions. After correcting school/program matching, the frozen v8 prompt achieved **77.04% accuracy** (95% post-bootstrap interval 75.58% to 78.45%), with **92.78% accuracy on confident calls at 32.08% coverage**. Its Brier score was 0.16168. The same model given a one-line request achieved 75.35% accuracy and Brier 0.17721. These results describe self-selected online posters, not calibrated personal probabilities for the general applicant population. The old test and an initial College Confidential sample were subsequently used for error analysis and cannot validate later prompts. On the separate 448-profile fresh comparison, V10 scored 77.52% versus 79.21% for frozen v8. The predeclared replacement rule selects the existing frozen v8. No result establishes a universal accuracy ceiling or rules out training-data contamination.

1. Research question

Can a prompt grounded in school admissions policies improve the accuracy and calibration of a generic chatbot response to an applicant profile? We compare decisions, probabilities, and the reliability of the subset called confident. High confident-call accuracy is meaningful only together with its coverage: it is not overall accuracy, nor a guarantee for a particular applicant.

The research prompt carries cards for 70 extensively researched colleges and 94 further schools, with University of California review rules, program-specific distinctions, reader-style evaluation guidance, and development-sample admit rates. The study tests the combined prompt; it does not separately establish the accuracy of every institutional fact in those cards.

2. Data and evaluation lineage

2.1 Original Reddit corpus

The archived r/collegeresults corpus contains 1,124 usable self-reported applicant profiles, 17,638 decision rows, and 15,751 scorable decisions. The original split allocated 700 profiles to development and 424 to testing. One test profile has no eligible scored decision, so the test metrics use 423 profiles and 6,129 decisions. A fixed 300-profile subset supported development comparisons.

Profiles were stripped of decision sections, outcome-revealing titles, commitment statements and retrospective commentary. The list of schools, rounds and programs was retained. The original collection used multiple audit rounds; those audits reduce rather than eliminate leakage risk. Source authors were hashed during collection, and the reported research does not reproduce individual profiles.

The original test was held out when v8 was frozen. Its errors were later analyzed to construct v9 and V10. Consequently, the corrected original-test results validate the historical v8 run only. Rescoring saved outputs after a matching correction is distinct from generating another prediction run. No V10 claim in this paper is based on the original test.

2.2 College Confidential

A separate corpus contains 1,408 profiles. An initial 500-profile sample supplies 2,342 scorable decisions in the archived v8 comparison. That sample was independent of v8 development but subsequently contributed to error analysis. It is not an untouched validation set for V10. The remaining College Confidential profiles were not used in this fresh-set comparison.

2.3 Reserved fresh Reddit set

The next comparison uses exactly the 448 IDs in `data/fresh/ids.txt`, distinct from the original development and test profile IDs. The collection notes report 6,667 decision rows and 5,975 scorable rows; the deterministic scorer's eligibility rule excludes 40 low-confidence rows, leaving 5,935 scored decisions from 446 profiles. All 448 profile output files were generated and validated. Only three profiles are dated after the stated model cutoff, so this set cannot provide a meaningful post-cutoff contamination control.

The recovery audit found that 438 of these profiles also exist in a separately collected admissions database corpus. That overlap is disclosed: freshness here means held out of this prompt's development and outcome analysis, not globally unknown to every project or necessarily absent from model pretraining. Prediction subprocesses receive neither that corpus nor retrieval tools.

3. Prediction and scoring methods

3.1 Isolated prediction runs

Each prediction runs through the configured Claude Code CLI using `claude-opus-5-5`, in an empty temporary directory with tools, MCP access, settings sources and session persistence disabled. The chancing predictor receives the frozen system prompt and one stripped profile. The lazy baseline receives a generic assistant system prompt and the same profile with a one-line request. Its answer is converted to per-school probabilities by a separate Sonnet session that sees the answer and school list, never outcomes. This converter is part of the baseline measurement procedure.

V10's exact bytes were frozen before this run. SHA-256:

```
e4df2aea9c5fa5f8395d5d3da3fee10dc61464e2ad5b631b3ecc3b7d369e0a0a
```

The run manifest records the ID-list hash, threshold, confidence band, model and selection rule. The subscription session limit had reset when the run resumed, but a weekly limit stopped it after 290 valid V10 profiles. The remaining 158 were resumed through the already configured Anthropic API key using the same Claude Code executable and

model. An API-backed health check confirmed `claude-opus-5-5`; no alternative model was substituted. This billing/authentication change is logged in `results/runner-fallback.json`. Frozen v8's existing fresh predictions were reused, not regenerated after seeing the comparison.

3.2 Outcome and metric rules

Accepted and deferred-then-accepted outcomes are positive; a waitlist with a stated later acceptance is positive. Rejected outcomes and other waitlists are negative. Unknown, withdrawn, pending and ineligible low-confidence rows are excluded according to the original protocol. Predictions are matched to canonical schools, then programs. A plain school row prefers a plain school prediction when both plain and program-specific predictions exist.

The operating point remains **Admit at p >= 0.42**, with **confident calls at p <= 0.12 or p >= 0.80**. The original default, p >= 0.5 with confidence bands 0.10/0.90, is also reported for the original test. Neither threshold nor confidence band was optimized on the fresh set.

Missing predictions count as incorrect in accuracy. For probability metrics they receive the protocol's p = 0.5 fallback, and they never count as confident. Expected profile IDs fix the denominator. Paired comparisons retain every decision row, including repeated school/program rows, and resample whole profiles rather than treating correlated decisions as independent. Reported intervals use 2,000 bootstrap resamples. Aggregate-score intervals use seed 11; paired differences use seed 3.

3.3 Runner repair

The previous runner treated an existing output file as completed even if the file contained a failed call or no predictions. Of the 448 fresh lazy-baseline outputs, 398 were valid, 49 existing files were invalid and one file was absent. The repair validates nonempty prediction arrays, school names and finite numeric probabilities in [0, 1], and rejects error records. Writes are atomic; replaced files are preserved under `attempts/`. Rate-limit failures stop the run rather than becoming predictions. Valid outputs are reused on resume.

The 50 affected baseline profiles are retried using the original model and baseline procedure. A nonempty response may still omit a school; that omission is scored as missing rather than silently excluded. Fixing execution failures does not change the prompt or choose outputs based on correctness.

3.4 Frozen replacement rule

Before V10 prediction began, the selection rule was recorded: replace frozen v8 only if V10's paired accuracy gain has a 95% post-bootstrap interval entirely above zero **and** V10's Brier score is no worse. Otherwise retain v8. This is a conservative product rule declared for this comparison, not a claim that it was part of the original October 15 protocol. No prompt edits, threshold searches or fresh-outcome-guided retries are permitted. After this comparison, the fresh set is spent; a later prompt needs another untouched test.

4. Corrected original-test results

Predictor	Accuracy	Confident accuracy	Coverage	Brier
Frozen v8	77.04%	92.78%	32.08%	0.16168
Lazy baseline	75.35%	86.84%	49.73%	0.17721
Feature logistic baseline	73.58%	89.65%	25.84%	0.17379

All three rows use the same 6,129 decisions. Missing school predictions are 1 for v8, 7 for the lazy baseline and 38 across five missing profile files for the logistic baseline; all count as incorrect. Frozen v8's confident-call accuracy interval is 91.22% to 94.13%; its coverage interval is 29.98% to 34.19%. The default operating point gives v8 76.41% accuracy, 93.71% confident accuracy and 21.52% coverage. Full metrics and paired comparisons are in Appendix A.

The prior manuscript reported 76.8% overall and 92.3% confident accuracy. Those are superseded, not additional experiments: the saved outputs are unchanged, and the school/program matching correction yields 77.04% and 92.78%. The old documents are retained in `paper/pre-correction/` and `results/pre-correction/`. The logistic baseline is also corrected to include 38 previously omitted decisions from absent profile output files; its old approximately 74.0% figure used an incomplete denominator.

Frozen v8 scores 72.42% on the original UC subset and 77.71% on other schools. Its after-cutoff score is 77.28%, versus 76.99% before the stated cutoff. These similar rates and the model's negative recall probe are limited diagnostics, not proof against memorization. The model's self-report cannot establish what was in training data.

On the historical College Confidential sample, frozen v8 scores 74.89% overall and 91.90% on confident calls at 34.80% coverage; the lazy baseline scores 73.65% overall and 89.13% confident accuracy at 41.25% coverage. These numbers are historical v8 evidence, separate from the V10 selection below.

5. Fresh-set V10 comparison

Prompt	Decisions	Accuracy	Confident accuracy	Confident coverage	Brier	Missing decisions
Frozen v8	5935	79.21%	94.21%	38.74%	0.14929	0
V10	5935	77.52%	95.07%	32.81%	0.14926	3
Lazy baseline	5935	77.94%	89.52%	51.78%	0.16050	5

Winner under the frozen rule: **Frozen v8**.

V10 minus frozen v8 accuracy: -1.68 percentage points (95% post-bootstrap interval -2.75 to -0.66). V10 minus frozen v8 Brier: -0.00003 (95% interval -0.00324 to +0.00296).

All 448 expected profile outputs passed validation; 446 profiles contribute 5,935 eligible decisions. Missing school predictions count as incorrect; Brier uses p=0.5 for missing values. No prompt edits followed outcome inspection. The fresh set is now spent and cannot serve as an untouched test of a later revision.

These results describe self-selected online posters. They do not validate personal admissions probabilities for the general applicant population.

V10 accuracy 95% interval: 76.04% to 78.99%; frozen v8: 77.78% to 80.57%. The paired interval, rather than overlap of these marginal intervals, determines the replacement decision.

6. Product recommendation

Retain and freeze **v8** for the product's research predictor. On the fresh set, v8 made 100 more correct calls than V10 out of 5,935 decisions: a 1.68-point lead (95% paired interval 0.66 to 2.75 points). V10's Brier difference is negligible and its interval includes zero. V10 therefore did not satisfy the predeclared replacement rule. This is evidence favoring v8 on this cohort, not proof that it is superior to every alternative.

The selected bytes, model and operating point are saved in `prompts/build/product.md`, `prompts/build/product.sha256` and `results/product-selection.json` once evaluation is complete. This freezes a research candidate. It does not deploy a public `/chances` route. The chance-page redesign remains a separate local preview for Leo's review.

Do not advertise confident-call accuracy as overall accuracy. Any public research claim should identify the online-poster population, print confident accuracy together with coverage, and distinguish a predicted probability from a validated individual probability. The proposed conversion from poster odds to general-population odds and an optional essay rubric were not independently validated by this experiment and are not included in the selected tested bytes.

7. What the exploratory analyses do and do not show

Earlier experiments compared prompt versions, ensembles, retrieval and fitted recalibration. Corrected development results appear in Appendix A. They show limited gains within the methods tested; they do not prove that all future methods must plateau at 77%.

The archived hindsight analyses chose an applicant's best cutoff after observing outcomes. Their roughly 87% fit describes particular school-ordering and single-cutoff model families. It is neither an upper bound on every predictor using profiles nor evidence that school ordering is already optimal. Essays, recommendations, school context and institutional priorities plausibly contain information absent from a self-report, but this experiment does not identify the amount of irreducible uncertainty.

The source research includes Common Data Sets, UC review documents and dashboards, the SFFA trial record and Opportunity Insights admissions research. These sources informed the prompt and exploratory population comparisons. Historical population differences and observational odds ratios do not justify an individual-level correction without further validation.

8. Limitations and research use

The data are self-selected, self-reported and incomplete. Stronger applicants and positive outcomes may be overrepresented. Ground truth comes from parsed text and includes judgment calls; later waitlist outcomes may be absent. Repeated users across sources may not be fully detectable. Residual leakage can remain in profile wording and application-list structure. The fresh set largely predates the stated model cutoff.

Only one primary predicting model was used in the final comparison. Historical calls and new calls were made at different times, so serving changes under the same model identifier cannot be ruled out. Reusing frozen v8 outputs compares the saved incumbent run to the new V10 run, not repeated samples of each. Bootstrap intervals reflect variation across sampled profiles, not full model-run variance, annotation error, or every source of uncertainty. The long research prompt also has material latency and inference cost.

There are no production database writes, migrations, emails or live UI changes in this research update. Individual source profiles and raw outputs remain local research artifacts. A later product integration and any general-population calibration require separate review.

9. Reproducibility

From the local research directory, `bash scoring/resume_fresh.sh` validates/resumes prediction files and regenerates metrics, tables and the manuscript. `python3 scoring/finish_research.py` performs deterministic evaluation only. `python3 results/make_tables.py > results/tables.md` regenerates the historical tables. `python3 paper/update.py` rebuilds this manuscript from the saved metrics. `python3 -m unittest discover -s scoring -p 'test_runner_io.py' -v` checks cache failures, quota handling, fixed denominators and paired duplicate rows.

The exact V10 prompt and ID-list hashes are in `results/fresh-evaluation-freeze.json`. `results/evaluation-provenance.json` records corpus and prediction hashes for the completed evaluation. Paths to raw inputs are local; preserving a script without the associated dataset does not by itself make a public release fully reproducible.

Appendix A. Corrected historical tables

Aristo Scholars (aristoscholars.com). Corrected scoring; historical tests apply to frozen v8 only.

Table 1. Held-out test (424 posts), operating point fixed on the benchmark: Admit at p >= 0.42, confident at p <= 0.12 or p >= 0.80

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
Aristo Scholars chanceme prompt (final)	6129	77.04	92.78	32.08	0.162	0.491	0.845	76.4	0.065
Lazy prompt, same model (B3)	6129	75.35	86.84	49.73	0.177	0.541	0.826	74.7	0.086
Stats-only logistic model (B2)	6129	73.58	89.65	25.84	0.174	0.522	0.814	73.6	0.023
Poster base rate per school (B1)	6129	73.39	89.54	14.67	0.186	0.553	0.793	73.2	0.038
All reject (B0)	6129	53.43	53.43	100.00	0.466	4.289	0.500	50.0	0.466

Table 1. Held-out test (424 posts), pre-registered: Admit at p >= 0.5, confident at p <= 0.10 or p >= 0.90

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
Aristo Scholars chanceme prompt (final)	6129	76.41	93.71	21.52	0.162	0.491	0.845	75.2	0.065
Lazy prompt, same model (B3)	6129	74.71	86.89	36.47	0.177	0.541	0.826	73.9	0.086
Stats-only logistic model (B2)	6129	74.40	92.05	16.84	0.174	0.522	0.814	74.0	0.023
Poster base rate per school (B1)	6129	70.35	95.00	4.24	0.186	0.553	0.793	69.2	0.038

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
All reject (B0)	6129	53.43	53.43	100.00	0.466	4.289	0.500	50.0	0.466

Missing decisions, included as incorrect: Aristo Scholars chanceme prompt (final): 1; Lazy prompt, same model (B3): 7; Stats-only logistic model (B2): 38; Poster base rate per school (B1): 0; All reject (B0): 0.

Table 2. Paired differences on the held-out test (bootstrap over posts, 2,000 resamples, 95% interval)

Comparison	Accuracy difference (points)	Brier difference
Final vs Lazy prompt, same model (B3)	+1.7 (+0.8 to +2.7)	-0.0155 (-0.0197 to -0.0115)
Final vs Stats-only logistic model (B2)	+3.5 (+2.2 to +4.8)	-0.0121 (-0.0175 to -0.0072)
Final vs Poster base rate per school (B1)	+3.7 (+2.5 to +4.9)	-0.0247 (-0.0302 to -0.0194)

Table 3. Final prompt by stratum (held-out test, operating point)

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
UC campuses	776	72.42	87.61	30.15	0.190	0.564	0.805	72.6	0.105
Non-UC	5353	77.71	93.48	32.36	0.158	0.481	0.850	76.7	0.059
Schools with a full card	4818	74.66	90.35	25.18	0.177	0.531	0.788	69.5	0.057
Other schools	1311	85.81	96.68	57.44	0.106	0.345	0.884	72.5	0.097
Posted before the model's training cutoff	4980	76.99	93.13	32.15	0.159	0.485	0.847	76.3	0.058
Posted after the training cutoff (2026-07 onward)	1149	77.28	91.23	31.77	0.173	0.519	0.836	76.4	0.098
Excluding rows with early-round list leakage risk	5932	77.11	92.82	32.62	0.161	0.490	0.847	76.6	0.067
Cycle 2021 (fall entry)	365	80.82	94.12	32.60	0.145	0.452	0.870	79.7	0.072
Cycle 2022 (fall entry)	418	72.49	93.33	28.71	0.185	0.542	0.813	72.8	0.099
Cycle 2023 (fall entry)	376	73.14	95.45	29.26	0.168	0.496	0.837	72.4	0.067
Cycle 2024 (fall entry)	419	75.18	90.00	28.64	0.173	0.526	0.814	74.4	0.077
Cycle 2025 (fall entry)	1270	77.32	92.12	34.96	0.157	0.482	0.843	76.9	0.043
Cycle 2026 (fall entry)	3281	77.78	92.88	32.09	0.160	0.488	0.851	76.9	0.073

Table 4. Waitlist sensitivity (waitlists excluded instead of counted as not admitted)

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
Final	5082	77.47	93.16	36.56	0.168	0.504	0.880	79.0	0.145
Lazy (B3)	5082	75.66	87.58	53.09	0.183	0.556	0.864	77.2	0.166

Table 5. Reliability of the final prompt (held-out test)

p bin	Decisions	Mean p	Observed admit rate
0.00-0.05	121	0.034	0.074
0.05-0.10	509	0.069	0.104
0.10-0.20	1136	0.145	0.172
0.20-0.35	1525	0.260	0.310
0.35-0.50	930	0.412	0.502
0.50-0.65	509	0.565	0.705
0.65-0.80	403	0.711	0.829
0.80-0.90	438	0.847	0.941
0.90-0.95	271	0.920	0.982
0.95-1.00	287	0.963	1.000

Table 6. Selective prediction: accuracy vs coverage (held-out test, final vs lazy)

Coverage (most confident share of decisions)	Final accuracy %	Lazy accuracy %
10%	98.20	92.97
20%	95.92	91.43
30%	93.85	90.75
40%	91.35	89.43
50%	89.13	86.78
60%	87.19	84.25
80%	82.42	80.52
100%	77.04	75.35

Table 7. Development history on the fixed 300-post benchmark (operating point)

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
v1 base method (100-post subset)	1391	75.92	86.07	57.80	0.186	0.564	0.865	76.3	0.172
v2 + recalibration table (100-post subset)	1254	74.72	90.76	38.84	0.168	0.517	0.827	74.4	0.072

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
v3 reader ratings + line	3878	76.95	90.86	41.77	0.166	0.501	0.848	76.8	0.086
v4 poster base rates + error-analysis rules	3878	76.66	92.18	39.92	0.166	0.498	0.853	76.5	0.094
v5 operating point + 94 short cards	3878	76.35	93.13	37.52	0.164	0.494	0.856	76.2	0.093
v6 + 392 reference cases (stopped at 114 posts)	1284	77.02	94.30	38.24	0.163	0.489	0.867	77.2	0.111
v7 + admissions-officer reading rules	3878	76.43	93.36	37.29	0.165	0.496	0.855	76.3	0.095
v8 + UC GPA-band table	3878	76.35	92.69	37.42	0.166	0.499	0.853	76.2	0.095
Lazy prompt (B3)	3878	75.04	87.69	49.02	0.178	0.539	0.834	74.9	0.099

Table 8. External validation on College Confidential (500 sampled applicants, independent of v8 development; later used for error analysis)

Predictor	Decisions	Accuracy %	Confident-call accuracy %	Confident coverage %	Brier	Log loss	AUC	Balanced acc %	ECE
Final	2342	74.89	91.90	34.80	0.172	0.514	0.827	72.7	0.114
Lazy (B3)	2342	73.65	89.13	41.25	0.188	0.561	0.808	71.8	0.133

Appendix B. Audit trail

`PROTOCOL.md` preserves the original rules and dated amendments. The pre-correction manuscript and tables remain available for comparison. `results/fresh-evaluation-freeze.json` predates this run; `results/fresh-evaluation.json` records the final metrics and paired comparison. `results/product-selection.json` records the selection outcome. All artifacts belong to Aristo Scholars, the product at aristoscholars.com; this research update is local and does not ship a live page.