MBA AI Engine Perception Index

Companion Document · Technical Validation, Source Audit, and Sensitivity

Publication revision 1.2.2 | September 2026 | Arcalea

ARCALÉA

arcalea.com/aeo-index/mba-aepi

MBA-AEPI | Scope clarification

September 14, 2026 | Supplement to publication revision 1.2.2

Full-time MBA focus

This release examines 25 tracked U.S. business schools using questions focused on full-time MBA education. Of 600 prompts, 498 explicitly specify full-time study; 102 use broader MBA or business-school wording. Full-time includes one-year and two-year options. Scoring identifies school mentions without independently verifying the MBA format discussed in each answer.

Conditions for interpreting the results

Results reflect this cohort and question mix. Studies targeting other MBA formats, specializations, applicant profiles, geographies, or school cohorts could produce different mention rates, rankings, and configuration differences. Those alternative outcomes have not been measured here. Full-time focus is a study condition; results should not be generalized to part-time, executive, or online MBAs. The 540-prompt sensitivity check addresses exclusion of named-school factual prompts, not robustness across MBA formats.

This clarification changes the description of the study scope, not its raw responses, scoring, numerical results, or statistical classifications. Original page numbering is retained on the following pages.

1. Purpose

This companion to the Methodology (publication revision 1.2.2) documents the empirical validation of the ChatGPT results: how the original optional-search ranking relates to an external prestige measure, what diverges, what the paired tests can and cannot detect, what the recovered sources reproduce, and whether the headline findings depend on the 60 named-school factual prompts.

> **WHAT CHANGED SINCE REVISION 1.2.1**
>
> The original raw baseline (1,200 responses), the 600-prompt registry with hashes, and both collection scripts were recovered and imported unchanged; both paired comparisons now reproduce from raw text. The registry audit found that all 60 Factual prompts explicitly name schools, so the aggregate index is a mixed discovery/lookup index and the former entity-free description is withdrawn. A post-hoc 540-prompt sensitivity analysis (Section 6) preserves the principal findings. Numerical results are unchanged.

2. Prompt Registry Audit

All 600 stored SHA-256 hashes match the UTF-8 prompt text, and prompt IDs and strata align with both raw collections. The 540 Generic, Constraints, and Persona prompts passed an automated name-token scan for target schools; that check is narrower than human validation of neutrality. All 60 Factual prompts (MBA-PROD-0541 to 0600) name schools: three question variants for each of 20 schools, 19 in the scored cohort and Emory Goizueta outside it. Consequently 57 prompts directly name tracked schools, three name an untracked school, and six tracked programs receive no direct factual question.

The retained 540 prompts also contain related and repeated wordings, so they should not be read as 540 independent query formulations. The factual questions are labeled Fact Verification & Program Data and are structured as school-specific lookups, consistent with purposeful inclusion; the original rationale for unequal coverage is not documented. The defensible reading is that the 600-prompt results are the original mixed research-task index, with factual-query behavior reported separately and the 540-prompt analysis disclosed as a sensitivity check. Aggregate rankings must not be presented as purely spontaneous discovery.

3. Prestige Correlation (original optional search)

The original analysis regressed M1-AGSO, a proportion bounded in [0,1], on U.S. News rank by OLS. That specification predicts negative mention rates at ranks 22–25 and produces an S-curve residual pattern that is a property of the functional form. The primary specification is rank-on-rank OLS: AI rank (1 = highest mention rate) regressed on U.S. News rank. Descriptive only; not recomputed for forced search.

Metric	Rank-space (AI rank ~ US News rank)	Score-space (M1 ~ US News rank, reference)
Spearman rho	0.973	−0.977
Pearson r	—	−0.941
R^2	0.944 (adj. 0.941)	0.886
Unexplained variance	5.6%	11.4%
p-value	< 0.001	< 0.001

The two specifications share data; the 5.6% and 11.4% figures describe different outcome scales and are not interchangeable. Neither is a measured independent AEO effect. The sensitivity analysis on 540

prompts returns identical values (Spearman 0.972843; R^2 0.943545) because the excluded Factual observations did not change any optional-search rank.

4. Residual Analysis: AI Rank Premium and Deficit

Rank residual = actual AI rank minus predicted AI rank. Negative = better rank than predicted (over-indexed); positive = worse (under-indexed). Positions relative to a fitted line within this cohort on one date; not evidence about merit or cause.

Over-indexed: Columbia Business School (−3.48; US News 7, AI rank 4), Tepper (−2.98), Foster (−1.94), Olin (−1.77), Stanford GSB (−1.68). **Under-indexed:** Johnson (+3.90; US News 16, AI rank 20), Smith (+3.14), Booth (+2.44), McCombs (+1.94), MIT Sloan (+1.40). MIT Sloan's earlier appearance as a top-3 over-indexed program used an inflated detection count and the score-space fit; neither holds under the corrected data.

> **ON READING RESIDUALS**
>
> The rank-on-rank specification does not share the score-space fit's negative-prediction defect, so every program's rank residual can be read directly and by magnitude. Under forced search the ordering changes (Columbia falls to a tie for rank 7 at 0.360); forced-search residuals are not reported.

5. Statistical Classification and Power

Because the same 600 prompts ran in both configurations, the two mention rates are paired: prompts where both configurations agree carry no directional information, and only discordant prompts do. The primary label is an exact two-sided McNemar test on discordant pairs with Benjamini-Hochberg correction across 25 programs at $\alpha = 0.05$. The legacy CI-overlap / five-point Option D rule is retained separately as a diagnostic (forced search: 12 suppressed, 13 Stable) and is not the headline. Statistical significance and practical materiality are distinct; the paired-interval materiality assessment specified in Amendment v1.9.1 has not been performed.

Original optional search. Five programs at nominal $p < 0.05$, all in the suppression direction (McCombs q 0.0505, Ross 0.0505, Tuck 0.0664, Johnson 0.0664, Wharton 0.1004); none survived correction. All 25 Stable. Stable means the difference did not reach the threshold in this sample; equivalence is a separate question the test leaves open. The zero-citation Persona stratum still shows measurable differences (Stanford +0.183); with no confirmed retrieval, that contrast cannot be attributed to live search, and more paired observations cannot recover the missing invocation evidence.

Forced search vs. the reused September 2 baseline. 13 Search-Suppressed (Harvard, Wharton, Columbia, Stanford GSB, Tuck, MIT Sloan, Kellogg, Haas, Yale SOM, Fuqua, Ross, Booth, Anderson), 3 Search-Amplified (Carey +0.030, Foster +0.028, Smith +0.012), 9 Stable. Largest negative difference: Columbia −0.230. Of the five nominal optional-search candidates, Wharton, Tuck, and Ross were significant here; McCombs and Johnson were Stable. These are measured differences between runs three days apart under an unpinned alias and differing output-token settings; a concurrent baseline in the next collection will reduce the timing confound; isolating retrieval itself also requires matched request settings, model-version controls, and retained invocation evidence.

Rk	Program	M1 forced	95% CI	M1 no-tool	ΔM1	Rk opt.	BH class	Legacy Option D
1	Wharton	0.447	[0.407, 0.487]	0.650	−0.203	2	Suppressed	Suppressed

Rk	Program	M1 forced	95% CI	M1 no-tool	ΔM1	Rk opt.	BH class	Legacy Option D
2	Kellogg	0.408	[0.370, 0.448]	0.598	−0.190	5	Suppressed	Suppressed
3	Harvard Business School	0.387	[0.349, 0.426]	0.605	−0.218	3	Suppressed	Suppressed
4	MIT Sloan	0.385	[0.347, 0.425]	0.573	−0.188	6	Suppressed	Suppressed
5	Stanford GSB	0.383	[0.345, 0.423]	0.600	−0.217	1	Suppressed	Suppressed
6	Haas	0.382	[0.344, 0.421]	0.495	−0.113	7	Suppressed	Suppressed
7	Booth	0.360	[0.323, 0.399]	0.442	−0.082	8	Suppressed	Suppressed
7	Columbia Business School	0.360	[0.323, 0.399]	0.590	−0.230	4	Suppressed	Suppressed
9	Stern	0.235	[0.203, 0.271]	0.250	−0.015	11	Stable	Stable
10	Yale SOM	0.188	[0.159, 0.222]	0.268	−0.080	10	Suppressed	Suppressed
11	Ross	0.148	[0.122, 0.179]	0.233	−0.085	12	Suppressed	Suppressed
12	Tuck	0.133	[0.108, 0.163]	0.325	−0.192	9	Suppressed	Suppressed
13	Darden	0.118	[0.095, 0.147]	0.122	−0.003	14	Stable	Stable
14	Fuqua	0.115	[0.092, 0.143]	0.190	−0.075	13	Suppressed	Suppressed
15	Johnson	0.093	[0.073, 0.119]	0.070	+0.023	20	Stable	Stable
16	Foster	0.088	[0.068, 0.114]	0.060	+0.028	18	Amplified	Stable
17	Anderson	0.085	[0.065, 0.110]	0.130	−0.045	15	Suppressed	Stable
18	Marshall	0.082	[0.062, 0.106]	0.078	+0.003	17	Stable	Stable
19	McCombs	0.080	[0.061, 0.104]	0.088	−0.008	19	Stable	Stable
20	Tepper	0.077	[0.058, 0.101]	0.097	−0.020	16	Stable	Stable
21	Carey	0.043	[0.030, 0.063]	0.013	+0.030	24	Amplified	Stable
22	Kenan-Flagler	0.023	[0.014, 0.039]	0.028	−0.005	21	Stable	Stable
22	Smith	0.023	[0.014, 0.039]	0.012	+0.012	25	Amplified	Stable
24	Olin	0.017	[0.009, 0.030]	0.020	−0.003	22	Stable	Stable
25	Mendoza	0.005	[0.002, 0.015]	0.018	−0.013	22	Stable	Stable

Forced search, all 25 programs, ranked on forced-search M1. Source: mba_aepi_index_scores_wave1b_chatgpt_v3.csv. Ranks 7 and 22 are ties (competition ranking). Suppressed = Search-Suppressed; Amplified = Search-Amplified.

6. Sensitivity Analysis: Excluding the Named-School Factual Prompts

A post-hoc analysis, not a new pre-specified index. It retains 540 prompt IDs in both arms (240 Generic, 180 Constraints, 120 Persona; effective weights 44.4%, 33.3%, 22.2%; no reweighting) and excludes the entire Factual stratum symmetrically. Detection is unchanged; Wilson intervals use n = 540; exact McNemar and BH are rerun within each comparison, with no joint correction across analyses and no materiality gate. The two analyses share responses and are not independent replications.

Measure	Original 600 prompts	Sensitivity 540 prompts
Optional-search statistical labels	25 Stable	Same 25 Stable
Forced-search statistical labels	13 suppressed / 3 amplified / 9 Stable	Same programs in all categories
Optional-search ranks	Reference	All 25 unchanged
Forced-search ranks	Reference	Spearman agreement 0.9985; top six unchanged
Optional-search U.S. News Spearman	0.972843	0.972843
Optional-search rank-space OLS R^2	0.943545	0.943545
Optional-search citation presence	83/600 (13.8%)	33/540 (6.1%)
Forced-search citation presence	583/600 (97.2%)	524/540 (97.0%)

6.1 Forced-search top six

Rank (both)	Program	M1, 600	M1, 540	Δ vs. no-tool, 600	Δ vs. no-tool, 540	BH class (both)
1	Wharton	0.447	0.491	−20.33 pp	−22.59 pp	Suppressed
2	Kellogg	0.408	0.448	−19.00 pp	−21.11 pp	Suppressed
3	Harvard Business School	0.387	0.424	−21.83 pp	−24.26 pp	Suppressed
4	MIT Sloan	0.385	0.422	−18.83 pp	−20.93 pp	Suppressed
5	Stanford GSB	0.383	0.420	−21.67 pp	−24.07 pp	Suppressed
6	Haas	0.382	0.419	−11.33 pp	−12.59 pp	Suppressed

Class labels are abbreviated: Suppressed = Search-Suppressed, Amplified = Search-Amplified.

Rates rise because the low-mention Factual stratum leaves the denominator; these are not increases over time. Columbia remains the largest negative difference (−25.56 pp on 540 prompts) and Carey the largest positive (+3.33 pp); Foster +2.96 pp and Smith +1.30 pp remain amplified. All optional-search p/q values are unchanged because the excluded observations contributed no discordant pairs; among forced-search tests only Foster (q 0.0305) and Smith (q 0.0260) changed, both still below 0.05.

6.2 Rank movements under forced search

Program	Rank, 600 prompts	Rank, 540 prompts
Tepper	20	18
McCombs	19	20
Kenan-Flagler	22	23

All other forced-search ranks are unchanged (minimum-rank tie convention). **Interpretation.** The check resolves whether the headline pattern depended on the named-school Factual prompts: in these saved

responses it did not. It does not validate factual accuracy, remove prompt dependence, establish human-validated detection, or remove the September 2–5 timing gap and request-setting differences. Files: sensitivity/optional_540_scores.csv, forced_540_scores.csv, comparison_600_vs_540.csv, summary.json; regenerate with analyze_discovery_sensitivity.py.

7. Recovered Sources and Reproducibility

Four files were imported unchanged from the local research folder and fingerprinted in chatgpt_source_manifest.json: the 1,200-response original capture file, the 600-prompt registry, and the two collection scripts. The forced-search capture file already in the repository is byte-identical to the local copy. Credentials, pilot files, and other-engine captures were excluded.

What reproduces from raw responses (read-only verifier, no API calls): exactly 600 NGMO and 600 AGSO original rows with unique prompt/mode pairs; 600 forced-search rows with unique IDs aligned to the registry; all mention rates and stratum counts; all four paired cells for each of 25 programs in both comparisons; exact McNemar p and BH q values including the 13 / 3 / 9 result; recorded citation counts and invocation flags; Wilson intervals; the original rank correlation; and all 600 registry hashes. Both corrected scoring scripts also regenerate every checked-in output within floating-point tolerance.

What is not established: detector precision and recall against human labels; the exact historical command-line arguments, SDK versions, or runtime; capture-time hash validation (the scripts did not perform it); the previously described every-ten-prompts checkpoint (not implemented; source comments implying it are superseded); or future model outputs. The original script records its requested model argument, not a server snapshot. The forced-search tool_invoked flag is set by either a web_search_call item or a url_citation annotation and is not retained as separate signals; all 600 are True, including 17 uncited responses. The original citation extractor deduplicates URLs while the forced extractor counts annotations, so mean citation counts are not identically defined.

Setting	Original dual-mode script	Forced-search script
API	client.responses.create	client.responses.create
Tools	web_search_preview for AGSO only	web_search_preview with tool_choice="required"
Output cap	max_output_tokens default 1500 (CLI-overridable; override not archived)	No explicit max_output_tokens
Timestamps	Sept 2, 2026, 16:59:24–18:49:55 UTC	Sept 5, 2026, 01:14:52–02:16:52 UTC
Hash check at capture	Not performed	Not performed
Write pattern	Batch CSV at completion	Batch CSV + JSONL at completion

8. Data Quality Attestation (recovered original file)

Check	Result
Rows	1,200 (600 NGMO + 600 AGSO), unique prompt/mode pairs
Recorded status 200	1,200 / 1,200
Recorded attempt = 1	1,200 / 1,200
Registry hashes	600 / 600 match
Identical TDK/LSS pairs	0 / 600
Empty responses	0

Check	Result
Truncated responses	Not independently established; completion-status metadata not retained
Entity universe coverage	25 / 25 programs detected

Response length across the 1,200 recovered original responses (characters): minimum 207, median 1588.5, mean 1775.3, maximum 5964. Status and attempt fields are harness-recorded, not independently preserved transport traces.

9. What This Validation Establishes

Confirmed

• The original optional-search AI rank is strongly associated with the recorded U.S. News rank (Spearman 0.973; rank-on-rank R^2 0.944), a descriptive fit statistic, not an independent AEO effect.

• Under forced search, 13 programs were named significantly less often and three more often than in the earlier no-tool baseline; nine did not meet the threshold. The same programs hold in every category when the 60 named-school factual prompts are excluded.

• Both paired comparisons, all p/q values, all registry hashes, and all mention counts reproduce from the recovered raw responses without API calls.

• All 60 Factual prompts name schools; the aggregate index is a mixed discovery/lookup index.

Open questions

• Detector accuracy against human labels, for any of the 25 patterns.

• The isolated causal contribution of retrieval: the forced-search comparison differs from its baseline in collection date, output-token setting, and possibly model version.

• Whether the original optional-search answers invoked search where no citations appeared.

• Practical materiality of each classification, via the paired-difference interval specified in Amendment v1.9.1.

• Consumer ChatGPT behavior, program quality, training-data contents, enrollment effects, and stability across engines and over time.

Next steps

• Pair a no-tool baseline with required search, collected close together, with explicit invocation logging and a pinned model snapshot.

• Human-code a validation sample to quantify detection precision and recall.

• If unprompted discovery becomes the primary outcome, label that as a separate methodological change rather than substituting it for the original index.

MBA-AEPI ChatGPT publication revision 1.2.2 | Companion document | September 2026. Companion to Methodology revision 1.2.2; incorporates SOURCE_AUDIT.md and sensitivity/REPORT.md.