Barkhausen AI

REPORT · BA-R-2026-02

AI visibility in Chinese study-abroad decisions: a pre-registered study protocol

The companion evidence review (BA-R-2026-01) established that a large and rising share of prospective students consult AI assistants when deciding where to study, and that, to this publication's knowledge as of July 2026, no measured account of which institutions and services those assistants actually surface has been published.

“Every leap begins as a crackle.”

DOCUMENT	BA-R-2026-02
SERIES	Report
DATE	2026
SUBJECTS	Market behavior; Measurement & statistics
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

The companion evidence review (BA-R-2026-01) established that a large and rising share of prospective students consult AI assistants when deciding where to study, and that, to this publication's knowledge as of July 2026, no measured account of which institutions and services those assistants actually surface has been published. This document is the study protocol for the first measured edition: it pre-registers, before data collection, the estimands, the cell design, the sample-size targets and interval methods, the measurement channels and their calibration, the pooling and multiplicity treatment, the refusal handling, the entity-naming policy, and the criteria under which any change will be claimed. The protocol discloses the statistical design layer in full and withholds the operational layer — concrete phrasings and tooling — per the disclosure floor of BA-C-4. Deviations in the eventual edition will be disclosed against this document. No publication date is promised; the edition follows the data.

IN BRIEF

Before running a study, serious research fields write down — in public — exactly what will be measured and how, so that the eventual results cannot be quietly reshaped to flatter whoever ran them. This document does that for a study of what AI assistants tell Chinese students researching study abroad: which questions will be sampled, on which assistants, how many times, over what periods, and by what statistical rules a finding will count. It also fixes two honesty policies in advance: universities are reported by name, while commercial intermediaries are reported as anonymized classes in the first edition; and any claim that visibility changed must clear a pre-stated statistical bar. The measured report will follow this protocol; where it must deviate, it will say so explicitly.

KEY FINDINGS

1. This protocol pre-registers the design of what is, to this publication's knowledge as of July 2026, the first measured study of AI visibility in education decisions; likewise to its knowledge, no measured, pre-registered study of institution or service visibility in AI answers to study-abroad questions had been published as of July 2026.
2. The estimand is Visibility Probability per BA-C-2: the probability that an entity is mentioned in an answer, defined over a phrasing distribution per information need and conditioned on engine, time window, and region, with the measurement channel recorded as a cell dimension per BA-C-3 — and with five pre-registered information-need classes covering destination, institution, program, intermediary, and logistics questions.
3. Each primary cell (information need × engine × channel × monthly window × region) is sized at $n \geq 96$ observations — twelve phrasings, eight draws each, spread across at least twelve distinct days — the worst-case sample size for a ±10-percentage-point 95% interval under independent draws; the reported hierarchical interval widens beyond that floor to reflect between-phrasing variance.
4. Answers will be collected on at least three engines spanning international and China-accessible systems, on the consumer interface as the primary channel; any API-collected cell is paired with a same-window consumer-interface calibration subsample, per the measurement-channel requirements of BA-C-3.
5. Entity naming is fixed in advance: universities and public institutions are reported by name; commercial intermediaries (agencies and pathway providers) are reported as anonymized classes in the first edition, with mention frequencies, concentration, and cross-engine overlap reported in aggregate — no evaluative statement about any named or unnamed provider will be made.
6. Any claim that visibility changed between windows must satisfy the Barkhausen Criterion of BA-C-3 — interval-based significance, sustainment across at least two consecutive monthly windows, exclusion of flagged engine changes, and Benjamini–Hochberg false-discovery-rate control at $q = 0.05$ across all monitored cells.

CONTENTS

1 Summary

The evidence review in this series (BA-R-2026-01) established the premise: a large and rising share of the audience for education decisions consults AI assistants, Chinese students among them, and what those assistants answer is consequential enough that students act on it. What the public record does not contain — anywhere, to this publication's knowledge, as of July 2026 — is a measured account of the answers themselves: which institutions and services AI assistants surface when asked the questions Chinese students actually ask, how often, how stably, and in what company.

This document is not that account. It is the account's *protocol*: a pre-registration, published before data collection, of exactly what the first measured edition will estimate and by what rules. Pre-registration is the discipline, standard in experimental fields [3] [4], of committing to design and analysis before seeing outcomes, so that results cannot be quietly reshaped after the fact — the failure mode this series' whitepaper documents in AI-visibility reporting at large (BA-W-2026-01). A field whose common practice is the retrofitted screenshot is precisely the field where measurement should begin by binding its own hands.

The protocol is written to the conventions of this series and claims design conformance with BA-C-3. The eventual edition will report in conformance with BA-C-5, using the metrics of BA-C-2, and every claim excerpted from it will satisfy BA-C-4. What this document deliberately does not contain is the operational layer — the concrete phrasing bank, collection tooling, and platform handling — per the disclosure floor stated in BA-C-4 §5: the design a statistician needs to judge validity is public; the machinery a competitor would need to replicate operations is not.

2 1. Objectives

Primary objective. Estimate the Visibility Probability (VP) — as defined in BA-C-2 §1–§2 — of entities relevant to Chinese study-abroad decisions, per information-need class, engine, measurement channel, and monthly window.

Secondary objectives. (a) Estimate Share of Voice within pre-registered comparator sets (BA-C-2 §3) and Discovery Depth along pre-registered constraint dimensions (BA-C-2 §4). (b) Characterize the market structure of AI answers in this domain: the concentration of mentions (share of all mentions captured by the most-mentioned entities), the effective number of distinct entities surfaced per need, and the overlap of recommendation sets across engines, using set overlap and rank-biased overlap (BA-C-2 §5.3). (c) Establish a baseline against which later editions can evaluate change under the Barkhausen Criterion.

Population of interest. The information needs of prospective students from mainland China researching study abroad — not the students themselves; this is a study of answers, not of people. No human subjects are involved; no personal data is collected.

3 2. Estimands and cells

The estimand is the one fixed in BA-C-2 §1: for entity B and information need k,

$$\mathrm{VP} = \Pr\big(B \text{ mentioned} \mid \text{prompt} \sim Q_k,\ e,\ t,\ g\big),$$

where Q_k is the phrasing distribution for need k, e the engine — which, per BA-C-2, includes the specific interface and version — t the monthly window, and g the region and locale controls. The measurement

channel c is the deployment surface through which e is observed and is recorded as part of the cell: a cell is one combination (k, e, c, t, g), per BA-C-3's cell definition.

Information-need classes (pre-registered, five). Each class contains multiple concrete needs; the classes are disclosed here, the concrete phrasings are not (§7):

1. **Destination selection** — comparing countries and systems ("where should I do a master's abroad").
2. **Institution selection** — comparing named universities within a destination, including rankings-and-reputation questions, the largest single question category in the international-student survey evidence reviewed in BA-R-2026-01.
3. **Program selection** — field- and degree-specific choices, including English-taught programs and admission requirements.
4. **Intermediary selection** — agencies, pathway providers, and application services.
5. **Application logistics** — tests, visas, timelines, costs, and scholarships, where institutions and services are mentioned incidentally rather than solicited.

Language and locale. Phrasings are drawn in Simplified Chinese as the primary language of Q_k, with an English subset per need where real usage warrants it; the edition will report the phrasing-language composition per need. Locale controls (accounts, network vantage, interface language) are fixed per engine and disclosed in the edition per BA-C-3's region requirements.

Entity classes. Detected entities are classified as: universities and public institutions; commercial intermediaries (agencies, pathway providers, application services); information platforms (rankings sites, forums, official portals). The naming policy per class is fixed in §6.

4 3. Engines and measurement channels

Engines. At least three engines are sampled, spanning the systems internationally dominant and those accessible and widely used from mainland China; the sampled set, with interface and version or observation date for every estimate, is disclosed in the edition per BA-C-2 §6. The set is not frozen in this protocol because engines launch, merge, and retire without notice; what is frozen is the disclosure obligation and the minimum count. Engines are reported separately, never pooled (BA-C-3, per-engine reporting).

Channels. The consumer interface — what a student actually sees — is the primary channel, and every claim in the edition about what users see rests on consumer-interface observations. Where any cell is collected through an official API for scale, that cell is paired with a same-window consumer-interface calibration subsample on the same needs, and the edition reports the observed divergence rather than asserting equivalence, per the measurement-channel requirements of BA-C-3. No cross-channel pooling occurs without disclosure. Where a widely used engine offers no compliant collection route at adequate scale, the edition reports it with reduced n and correspondingly wider intervals rather than silently dropping it — a smaller honest cell outranks a missing one.

Session controls. Observations use stateless sessions: no conversation history, personalization and memory disabled or absent, one observation per session. The edition states the mechanism per engine (BA-C-3, region and personalization).

5 4. Sample size, windows, and collection discipline

Per-cell target. Primary cells are sized at $n \geq 96$ observations: twelve distinct phrasings per need, eight draws per phrasing, spread across at least twelve distinct days within the monthly window (Figure 1). At worst case ($p = 0.5$), $n = 96$ yields a 95% interval half-width of ±10 percentage points under independent draws (BA-C-2 §2) — a floor, not a promise: because the draws are clustered within twelve phrasings and the estimand is an expectation over the phrasing distribution, the realized hierarchical interval is generally wider, and the edition reports that wider interval; most cells will sit nearer the boundaries, where the mandated Wilson or Clopper–Pearson intervals (BA-C-3) are narrower and remain valid. Cells that cannot reach the target in a window — manual-collection engines among them — are reported at their achieved n with their achieved interval, marked below-target, and never pooled into compliant cells to disguise the shortfall.

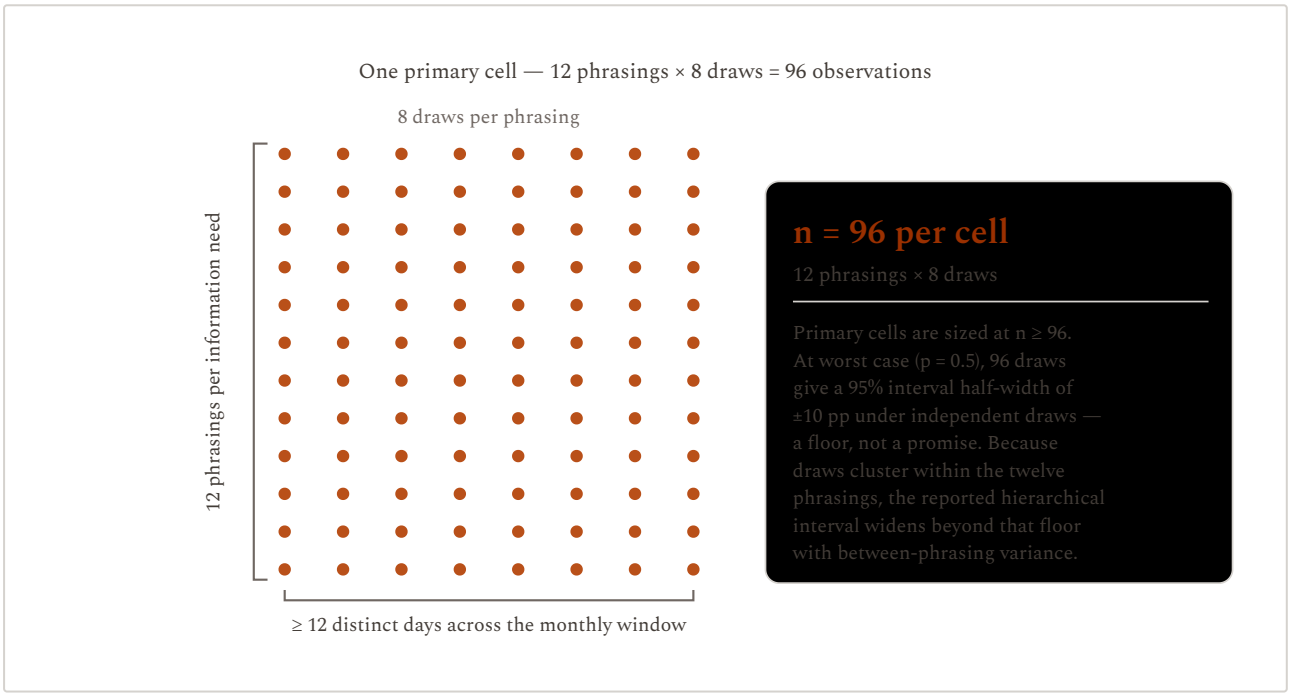

Figure 1. The design of one primary cell. Twelve phrasings per information need (rows) times eight draws per phrasing (columns) gives n = 96 observations, spread across at least twelve distinct days of the monthly window. At worst case (p = 0.5), n = 96 is the floor for a ±10-percentage-point 95% interval under independent draws; because draws cluster within phrasings, the reported hierarchical interval widens beyond that floor with between-phrasing variance.

Barkhausen AI · BA-R-2026-02

Why not fewer. The public evidence this series rests on shows same-need paraphrases replacing much of a recommendation set [1] and identical prompts drifting day to day [2]; twelve phrasings address the first, twelve collection days the second, and eight repeats per phrasing the request-to-request noise floor. The design follows the clustering requirements of BA-C-3: draws are spaced so that the effective sample size approaches the nominal count, and the hierarchical model below reports uncertainty that reflects the remaining correlation.

Windows. The unit window is one calendar month. The baseline edition reports at least one complete window per cell; change claims require at least two consecutive windows (§8).

Refusals and availability. Refusals are recorded as availability observations and reported per cell, with visibility estimates marked conditional where refusal rates are material (BA-C-3, refusals).

6 5. Analysis plan

Estimation. Within each cell, phrasing-level proportions are combined by Beta–binomial partial pooling — phrasings as units within the need, shrinkage governed by per-phrasing sample size — reported as posterior means with 95% credible intervals (BA-C-3, pooling); the credible interval, bounded within $[0, 1]$ by construction, serves as the boundary-valid interval for pooled estimates. Raw proportions, where quoted, carry Wilson or Clopper–Pearson intervals near the boundaries. Every published figure carries the full disclosure set of BA-C-4.

Detection. Mentions are detected by lexicon-and-alias matching over answer text, with entity disambiguation for confusable names. Before publication, the detector is validated against a human-labeled sample of at least 200 answers spanning engines and languages, and the edition publishes its precision and recall on that sample (BA-C-2 §6). Where a conclusion would rest on a difference smaller than the detector's error rates plausibly explain, the edition says so and does not draw it.

Market structure. Concentration is reported as the share of all mentions captured by the top-k entities per need class (k = 1, 3, 5) and as the effective number of entities (inverse Simpson index over mention shares). Cross-engine agreement is reported as set overlap (Jaccard) and rank-biased overlap with the persistence parameter disclosed (BA-C-2 §5.3).

Engine-change monitoring. Per-cell series are monitored with online change-point detection; broad, synchronous shifts are flagged as likely engine changes and reported as such, not as entity movement (BA-C-3, detecting engine change).

7 6. Entity-naming policy (pre-registered)

Fixed before any data exists, so that the naming decision cannot respond to what the data shows:

- **Universities and public institutions are reported by name.** Their appearance in AI answers is a fact about a public institution's public visibility, of the same kind this series' censuses report.
- **Commercial intermediaries are reported as anonymized classes in the first edition** — "Provider A–N," with class-level mention frequencies, concentration, and stability. This publication does not review, rank, or recommend vendors; a named mention-frequency table for commercial providers — unlike the public-institution reporting above, which records a public body's public visibility — would function as a vendor ranking regardless of framing, and the market-structure findings this study exists to establish do not require the names. The anonymization map is retained internally and applied consistently across editions so that longitudinal statements remain meaningful.
- **No evaluative statement** — quality, trustworthiness, desirability — is made about any entity, named or anonymized, in any edition. Mention frequency is reported as an observed property of engine output, with the explicit caveat that being mentioned is neither an endorsement by the engine nor by this publication.

8 7. What is published and what is withheld

Published with the edition: all aggregate estimates with their full BA-C-4 disclosure sets; the phrasing count, language composition, and coverage strategy per need; detector validation results; refusal rates; the aggregate dataset (CSV) under CC BY 4.0; and a completed BA-C-5 checklist, including the statement its commercial-interest item (BA-C-5 item 25) calls for. Withheld: the concrete phrasing bank, raw answer

transcripts, collection tooling, and platform-specific handling — the operational layer whose withholding BA-C-4 §5 anticipates. This is the standing disclosure grade of this series and does not vary by result.

9 8. Change claims and future editions

The baseline edition makes no change claims. Any later claim that an entity's or class's visibility changed must satisfy the Barkhausen Criterion as formalized in BA-C-3: interval-based significance between properly estimated cells; sustainment across at least $K = 2$ consecutive monthly windows (the value of K used is disclosed with the claim); non-coincidence with a flagged engine change, or re-basing after it; and multiplicity control by the Benjamini–Hochberg false-discovery-rate procedure at $q = 0.05$ [5] across all cells monitored in the comparison, with the cell count disclosed.

10 9. Deviations

The measured edition will carry a deviations section reconciling itself against this protocol, item by item, in the registered-report discipline [3]: what was executed as pre-registered, what deviated, why, and with what consequence for interpretation. A deviation disclosed is a limitation; a deviation discovered is a failure. Amendments to this protocol before data collection, should any be needed, are published as a new version of this document under the series' versioning rules, never edited silently.

11 Limitations

A protocol constrains analysis; it does not guarantee execution. The design commits to sample sizes and calibration subsamples whose feasibility differs by engine — manual-collection engines in particular may repeatedly land below target, and the edition will show exactly where.

The engine set is a moving target. Systems available and dominant at protocol time may be replaced by edition time; the protocol therefore binds disclosure obligations and minimum counts rather than a frozen engine list, and that is a real, acknowledged looseness — a fully frozen design would be obsolete before its first window closed. These results, when they come, will describe the engines as sampled during their stated windows; engines change without notice, and results should be assumed perishable.

The anonymization of commercial intermediaries trades citability for safety: readers wanting named provider tables will not find them here, in any edition, and analyses that require names — partner-level attribution, for example — are out of this study's scope by design.

Finally, self-publication of a protocol is registration by timestamp, not by registry: no independent pre-registration registry exists for this field. The protocol's force is reputational and checkable — the edition can be audited against this public, dated document — which is weaker than institutional registration and stronger than the field's current norm, which is nothing.

REFERENCES

1. Jack, Lehman, Maloney, Xu; arXiv:2605.27440. *Paraphrase Brittleness in Production Retrieval-Augmented Commercial Recommendation: Reproducibility Below the Rerun-Stability Baseline* (2026). https://arxiv.org/abs/2605.27440 Accessed 2026-07-09. [archived]

2. Schulte, Bleeker, Kaufmann; arXiv:2604.07585. *Don't Measure Once: Measuring Visibility in AI Search (GEO)* (2026). https://arxiv.org/abs/2604.07585 Accessed 2026-07-09. [archived]

3. Chambers, C. D.; Cortex 49(3), 609–610. *Registered Reports: A new publishing initiative at Cortex* (2013). https://doi.org/10.1016/j.cortex.2012.12.016 Accessed 2026-07-09.

4. Nosek, B. A.; Ebersole, C. R.; DeHaven, A. C.; Mellor, D. T.; Proceedings of the National Academy of Sciences 115(11), 2600–2606. *The preregistration revolution* (2018). https://doi.org/10.1073/pnas.1708274114 Accessed 2026-07-09.

5. Benjamini, Y.; Hochberg, Y.; Journal of the Royal Statistical Society, Series B 57(1), 289–300. *Controlling the false discovery rate: a practical and powerful approach to multiple testing* (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x Accessed 2026-07-09.

HOW TO CITE

Barkhausen AI (2026). AI visibility in Chinese study-abroad decisions: a pre-registered study protocol. https://barkhausen.ai/research/study-protocol-china-study-abroad/

Barkhausen AI · Est. MMXXVI · *Every leap begins as a crackle.*