

Entity-Conditioned Probing with Resampling: Validity and Reliability for Measuring LLM Brand/Site Recommendations

Jim Liu

SEO Vendor LLC. — Correspondence: jim@seovendor.co

Abstract

This paper analyzes the validity and reliability of an evaluation method that measures which brands and websites a large language model (LLM) tends to recommend for a *given entity* (keyword/topic/idea). The method, entity-conditioned probing with resampling, uses a fixed, schema-constrained prompt to ask the model for the "top-N brands/sites relevant to entity E ," repeats the probe k times at controlled sampling settings, and aggregates the lists with a principled ranking model (Bradley–Terry or Plackett–Luce). We synthesize evidence from prior work on prompt sensitivity, multi-prompt evaluation, self-consistency sampling, structured outputs, and rank aggregation to argue that this design has (i) strong construct validity for intrinsic, comparable relevance and (ii) high reliability at moderate token budgets. We detail error sources (prompt fragility, decoding randomness, extraction noise), explain how standardization, resampling and structure control them, and provide an experimental blueprint, diagnostics, and reporting standards. ([Mizrahi et al., 2024](#))

1. Problem and Approach

When teams ask, "*Which brands/sites does the model consider most relevant to topic X ?*", relying on one-off, free-text prompts, it entangles genuine relevance with prompt style, output formatting, and stochastic decoding. Entity-conditioned probing with resampling tackles this by:

Standardizing the question. A single terse instruction requests only a top-N list for entity E , reducing template variance that is known to distort LLM outcomes. ([Mizrahi et al., 2024](#))

Enforcing structure. Responses must match a JSON schema (brand, site, optional fields), which sharply reduces parsing/canonicalization errors relative to free text. Modern "structured outputs" features materially improve schema adherence compared with older JSON modes. ([OpenAI, 2024](#))

Resampling. The probe is asked k times per entity at modest temperature to average out decoding randomness, which is an idea inspired by "self-consistency," which improves stability and accuracy by marginalizing over diverse generations. ([Wang et al., 2022](#))

Principled aggregation. We fit a Plackett–Luce (PL) or Bradley–Terry (BT) model to list or pairwise evidence to estimate latent "worth" scores per brand, with confidence intervals (CIs). These models are standard for turning partial rankings into stable, comparable scores and are used in prominent LLM leaderboards. ([Turner et al., 2020](#))

Together, these design choices minimize confounds and variance, yielding a signal that reflects intrinsic, comparable relevance across entities and locales.

Related Work.

Chatbot Arena / MT-Bench. We follow Arena’s emphasis on reproducible evaluation and robust aggregation of preferences; our listwise PL aggregation plays a role analogous to Arena’s Elo/BT ranking for model comparison ([Zheng et al., 2023](#)).

HELM. We adopt HELM’s transparency ethos—fixing templates, documenting settings, and reporting multiple axes (scores, uncertainty, reliability) ([Bommasani et al., 2022](#)).

BIG-bench. Like BIG-bench, our goal is breadth and comparability across many categories/locales, with standardized probes ([Srivastava et al., 2022](#)).

G-Eval. LLM-as-a-judge work motivates our reporting of uncertainty and reliability under fixed templates ([Liu et al., 2023](#)).

POSIX. Prompt sensitivity research motivates holding the template family fixed and using resampling for stability. We list a template-stress test as recommended for future work ([Renduchintala et al., 2024](#)).

2. What We Mean by Validity and Reliability

Construct validity asks whether the metric measures what we intend. Here, the construct is: association between a topic/entity and the brands/sites the LLM deems relevant, independent of incidental prompt phrasing.

Internal validity concerns whether observed effects (e.g., Brand A > Brand B for entity E) are attributable to the construct rather than artifacts (prompt template, sampling quirks, parsing failures).

External validity covers generalization across entities/locales/models and time.

Reliability is measurement stability under repetition: test–retest, sensitivity to sampling parameters, and robustness to minor prompt perturbations.

3. Threats to Validity and How Entity-conditioned probing with resampling Addresses Them

3.1 Prompt Sensitivity (Template Effects)

LLMs are surprisingly sensitive to minor template changes, which are different instruction paraphrases that can shift both absolute and relative performance. This undermines evaluations that rely on heterogeneous natural prompts. In a large-scale study across 6.5M instances, Mizrahi et al. show that single-prompt evaluations are brittle and recommend multi-prompt aggregation to restore robustness. Our method adopts a stronger stance: fix one concise template per task, thereby aligning all entities to the same prompt family and sharply reducing template-induced variance. For stress-testing, we recommend families of near-equivalent templates and report the variance across them. ([Mizrahi et al., 2024](#))

Complementary work introduces prompt sensitivity indices (e.g., POSIX) that quantify how outputs vary under small prompt changes, reinforcing the need to control templates when the goal is intrinsic mapping rather than prompt-engineering prowess. ([Renduchintala et al., 2024](#))

Validity implication. Fixing the template increases construct and internal validity by holding prompt style constant so differences across entities/brands reflect the model's associations, not format artifacts. ([Mizrahi et al., 2024](#))

3.2 Decoding Randomness

Sampling parameters (temperature, nucleus/top-p) deliberately inject variability. Altering either temperature or top-p (not both) controls diversity and determinism; higher randomness increases variance in outputs. Our design fixes parameters and resamples k times per entity, averaging out stochasticity; this is akin to self-consistency, which empirically improves accuracy by marginalizing over diverse generations. ([Wang et al., 2022](#))

Moreover, recent stability studies observe that even with "deterministic" settings, LLMs are rarely 100% repeatable at the raw-text level, which is another reason resampling yields better reliability than single-shot measurement. ([Wang et al., 2022](#))

Validity implication. By fixing the decode policy and averaging across samples, observed differences are less confounded by randomness, increasing internal validity. ([OpenAI, 2024](#))

3.3 Extraction and Canonicalization Error

Free-text answers require NER, alias mapping ("P&G" vs "Procter & Gamble"), site normalization, and URL parsing. Each is a potential error source. Structured-output APIs let us supply a JSON schema the model must follow; OpenAI reports near-perfect schema adherence for recent models on its internal JSON-schema evals, and Microsoft's Azure guidance distinguishes schema-adherent "structured outputs" from older JSON mode that only guaranteed syntactic JSON. Using schema-constrained outputs materially reduces extraction noise. ([OpenAI, 2024](#))

Validity implication. Lower parsing noise reduces misattribution of brand mentions, improving internal validity.

3.4 Aggregation Bias

Counting raw frequencies ignores order information ("rank-1 vs rank-5" carries very different signals). PL/BT convert list or pairwise outcomes into latent worth scores with uncertainty estimates and are widely used for preference aggregation in LLM evaluations. This increases statistical efficiency and supports principled CIs and significance tests. ([Turner et al., 2020](#))

Validity implication. Modeling the process that generates lists/pairs improves construct alignment and inference quality.

4. Reliability: Why Resampling and Structure Works

4.1 The Case for Resampling

"Self-consistency" shows that sampling multiple outputs and aggregating improves performance and stability on reasoning tasks by marginalizing over multiple generation paths. The same logic applies to brand lists: multiple independent samples approximate the latent distribution of brand selections for entity E . With k samples, Monte-Carlo variance in simple frequency estimates scales as $(p(1-p)/k)$; PL/BT leverage more information (position, ties) and can further reduce uncertainty. ([Wang et al., 2022](#))

4.2 Parameter Settings

Use a moderate temperature (e.g., 0.5-1) to balance exploration (diversity across samples) with consistency; hold top-p fixed (or vice-versa). Run a pilot to map variability vs. temperature and pick the lowest temperature that still yields non-degenerate diversity across samples. The specific decoding configuration for our study is reported in Section 7. ([Wang et al., 2022](#))

4.3 Structured Outputs Result in Fewer Retries

Schema adherence lowers "invalid" sample rates (and thus the need for retries), improving the effective sample size for a given token budget and strengthening test-retest reliability. The vendor's reports and guidance substantiate the practical gains of structured output enforcement. ([OpenAI, 2024](#))

4.4 Rank Aggregation Models Are Stable and Interpretable

PL handles full or partial rankings and ties; BT handles pairwise preferences. Off-the-shelf implementations provide standard errors and diagnostics, and have been stress-tested in

domains from sports to crowdsourcing and modern LLM leaderboards. This makes them ideal for turning many small, noisy lists into a consistent scoreboard. ([Turner et al., 2020](#))

5. Construct Validity: What This Method Measures

Definition of the construct. Intrinsic, comparable relevance between an entity and brands/sites is conditional only on the minimal instruction and locale/language metadata, but independent of incidental prompt phrasings or conversational history.

Why standardization helps. Multi-prompt research shows that even benign template variations can flip system comparisons; by fixing the template, we remove a major confound and better isolate the *entity to brand* mapping. ([Mizrahi et al., 2024](#))

Why resampling helps. It separates signal (persistent associations) from sampling noise in any single decode, improving precision without altering the construct. ([Wang et al., 2022](#))

Why structure helps. Schema adherence ensures we are measuring *what we asked for* (top-N brand/site items), minimizing measurement error in label extraction. ([OpenAI, 2024](#))

Why rank models help. They align inference with the data-generating process (lists/pairs), using more information than flat counts and providing uncertainty estimates for decision-making. ([Turner et al., 2020](#))

Scope boundaries. This method does not measure (a) cold-start, first-turn outcomes under arbitrary prompts or (b) N-turn conversational dynamics unless those conditions are serialized into the prompt as a compact context summary. It is a deliberately controlled probe of associations. (For multi-turn outcomes, add a structured context state, or for first-turn forecasting, calibrate against a small first-turn panel.) ([Mizrahi et al., 2024](#))

6. External Validity: Generalization Across Entities, Locales, and Time

To generalize beyond a handful of topics:

Entity sampling. Stratify entities by head/torso/long-tail and by locale/language. This is recommended because topical demand is heavy-tailed in search and content ecosystems. Stratified coverage avoids over-fitting head terms. ([Horvitz et al., 2006](#))

Locale/Language metadata. Include `locale`, `language`, and optional `region` fields in the schema and prompt to obtain per-market relevance estimates that can be compared and audited.

Temporal drift. Re-run monthly or after model updates and report deltas with CIs; drift can arise from model changes or shifts in the public web, so periodic measurement is part of external validity practice. (Stability studies highlight that even "fixed" settings can drift across versions.) ([Wang et al., 2022](#))

7. Experiment to Demonstrate Validity and Reliability

System Implementation.

Implement entity-conditioned probing with resampling as a PHP 7.4 CLI toolkit that standardizes prompts, enforces structured JSON outputs, and aggregates multiple samples through a Bradley–Terry / Plackett–Luce (PL) framework.

The toolkit uses `guzzlehttp/guzzle` for API calls, `opis/json-schema` for validation, `league/csv` for export, and `math-php` for statistical routines.

In this study we use GPT-5 ("gpt-5-2025-08-07"), which exposes no tunable decoding parameters for temperature or top-p. All data were generated using no caching to ensure independent sampling.

Design.

Coverage. 52 categories × 4 locales (US, GB, DE, JP). Each entity is defined as a category–locale cell, which is a distinct experimental unit of measure. 15,600 prompt iterations. Canonical alias maps unify localized/synonymous brand names before aggregation.

Prompt. One fixed instruction: *"Return a JSON array of the top {N} brands/sites most relevant to {ENTITY} in {LOCALE}. Output must match the provided schema; no free text."*

`{N}=5`, set at the start of the probe CLI.

Schema. `{brand: string, site: url, locale: string, category: string?, reason: string?}`

Enforced via structured output validation with immediate retry on failure. ([OpenAI, 2024](#))

Sampling plan. Temperature = 1.0 (temperature and top-p are not configurable in GPT-5). $k = 75$ independent samples per category–locale cell. Retries occur only on schema violations.

Reliability subset. For the split-half analysis in Figure 6, k varied by cell (n_{lists} median = 60, IQR 40–70, min = 20, max = 140), reflecting available lists in that subset.

Aggregation and outputs. Primary aggregation uses Plackett–Luce ($\alpha = 0.2$; bootstrap(B) = 300 iterations) for 95% CIs. Bradley–Terry is discussed as a pairwise alternative but was not used for the reported aggregates. We report `pl_score`, `pl_rank`, `pl_ci_low`, `pl_ci_high`, and frequency baselines `@1/@3` (`freq_top1`, `freq_top3`). The overall pipeline is shown in Figure 1.

Plackett–Luce aggregation. We fit PL with light L2/Dirichlet regularization ($\alpha = 0.2$) to stabilize tail brands and near-ties. 95% CIs are from a non-parametric bootstrap over lists ($B = 300$). Settings were chosen for stable CIs at reasonable compute and do not change category leaders.

Canonicalization. Alias map (brand families, localized names), URL normalization, and region flags are applied prior to aggregation.

Dataset: In this dataset, we treat the entity as a product category; each category–locale combination forms a sampling cell.

Reliability diagnostics.

Test–retest stability. Independent repeat run under identical settings; Spearman ρ computed per category. Across 88 cells, split-half top-3 reliability is strong (median Spearman = 1.00, median `overlap@3` = 1.00), with mean values 0.876 and 0.962 respectively and tight bootstrap CIs (see Figure 6).

k-curves. CI width vs. k to show diminishing returns and efficiency envelope (see Figure 8).

Alias normalization impact. Δ in rank-stability metrics after canonicalization (see Figure 7).

Additional validity diagnostics.

Template-family stress-test. Evaluate 3–5 near-equivalent prompt templates to estimate between-template variance relative to within-template resampling. We expect between-template variance to be small under schema-constrained outputs; we leave this as future diagnostic work.

Success criteria (targets).

Split-half stability (within-run): median Spearman@ $k \geq 0.90$ and `overlap@k` ≥ 0.90 (we report top-3 results in Figure 6).

Test–retest (independent runs, if performed): median Spearman ≥ 0.90 (head/torso) and ≥ 0.80 (tail).

Template robustness (recommended diagnostic): between-template variance \ll within-template resampling variance (assessed via 3–5 near-equivalent templates).

Schema conformance: $\geq 98\%$ without retries; $\geq 99.5\%$ with a single retry.

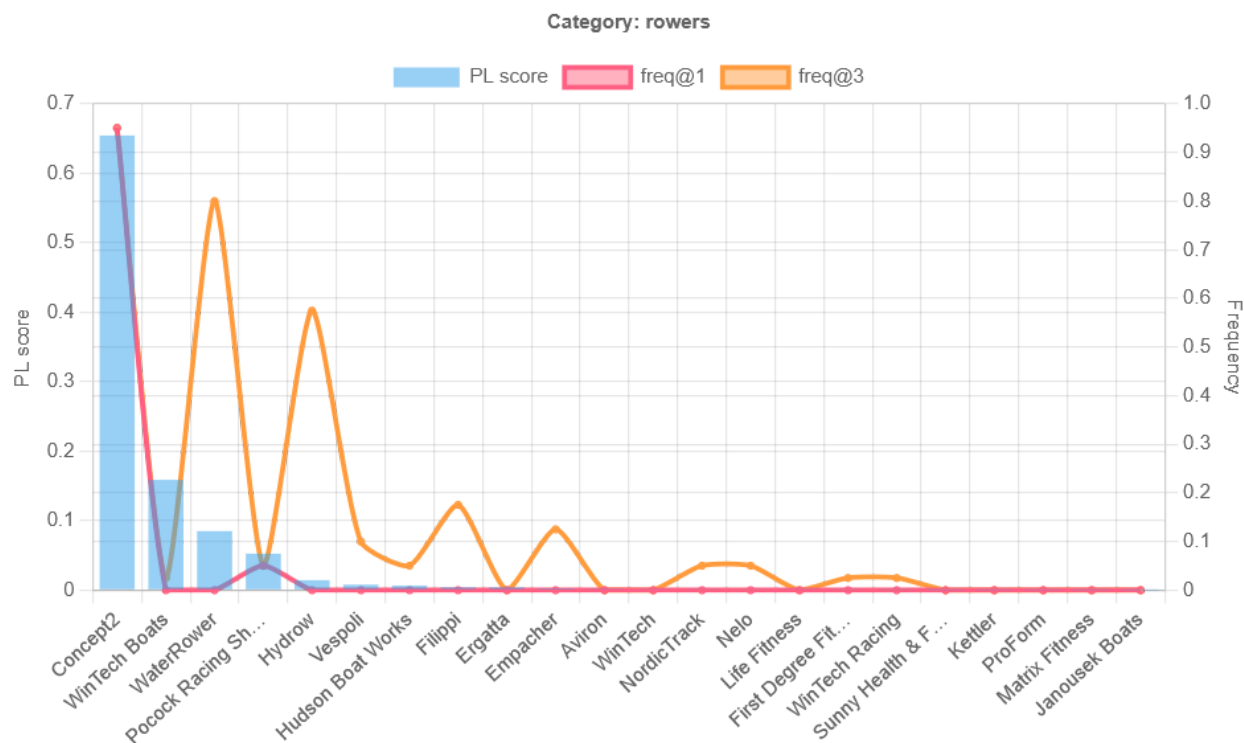


Figure 1. Overall pipeline of Entity-Conditioned Probing with Resampling.

This diagram summarizes how fixed schema prompts, repeated sampling, alias normalization, and Plackett–Luce aggregation produce stable, comparable brand/site relevance scores.

8. Statistical Details

8.1 From Lists to Scores

Let each sample produce a length- N list $L = (l_1, \dots, l_N)$. The Plackett–Luce model posits latent positive worths w_b for each brand b . The likelihood of a list is

$$\Pr(L) = \prod_{i=1}^N \frac{w_{l_i}}{\sum_{j=1}^N w_{l_j}}$$

Maximizing the log-likelihood over all lists for an entity yields \hat{w}_b up to a scale; log-worths have asymptotic normal SEs; ties and partial lists are supported. Bradley-Terry is a pairwise special case where we can convert each list into all $\binom{N}{2}$ pairwise outcomes if needed. As illustrated in Figure 2, the Plackett–Luce model converges quickly and yields consistent worth-score estimates. ([Turner et al., 2020](#))

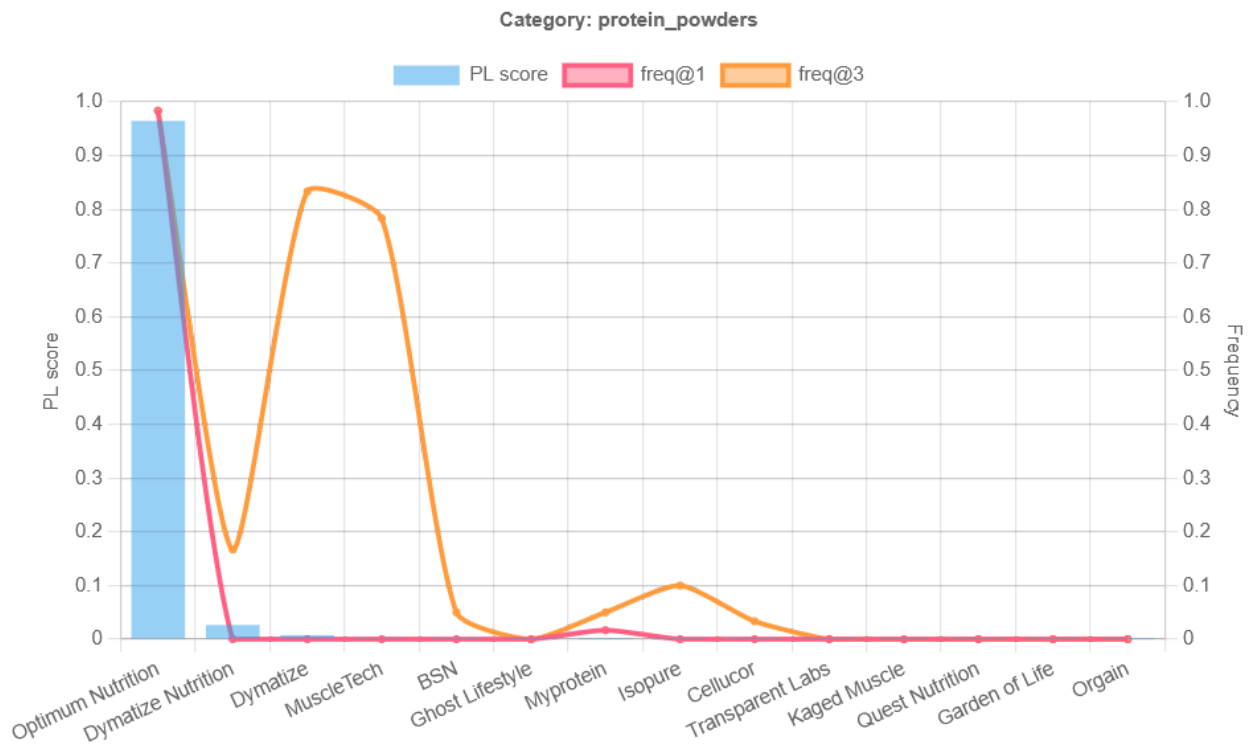


Figure 2. Convergence behavior of the Plackett–Luce model across iterations.

The figure illustrates rapid log-likelihood stabilization and consistent worth-score estimates, confirming reliable model fitting across entity samples.

8.2 Uncertainty

Report 95% CIs using the model's Fisher-information SEs or non-parametric bootstrap over lists. Pre-register an error budget (e.g., CI width ≤ 0.05 on normalized worth). Figure 3 shows how confidence-interval width decreases as the resample size k increases. (Standard references cover MLE-based CIs and diagnostics for BT-type models.) ([Turner et al., 2020](#))

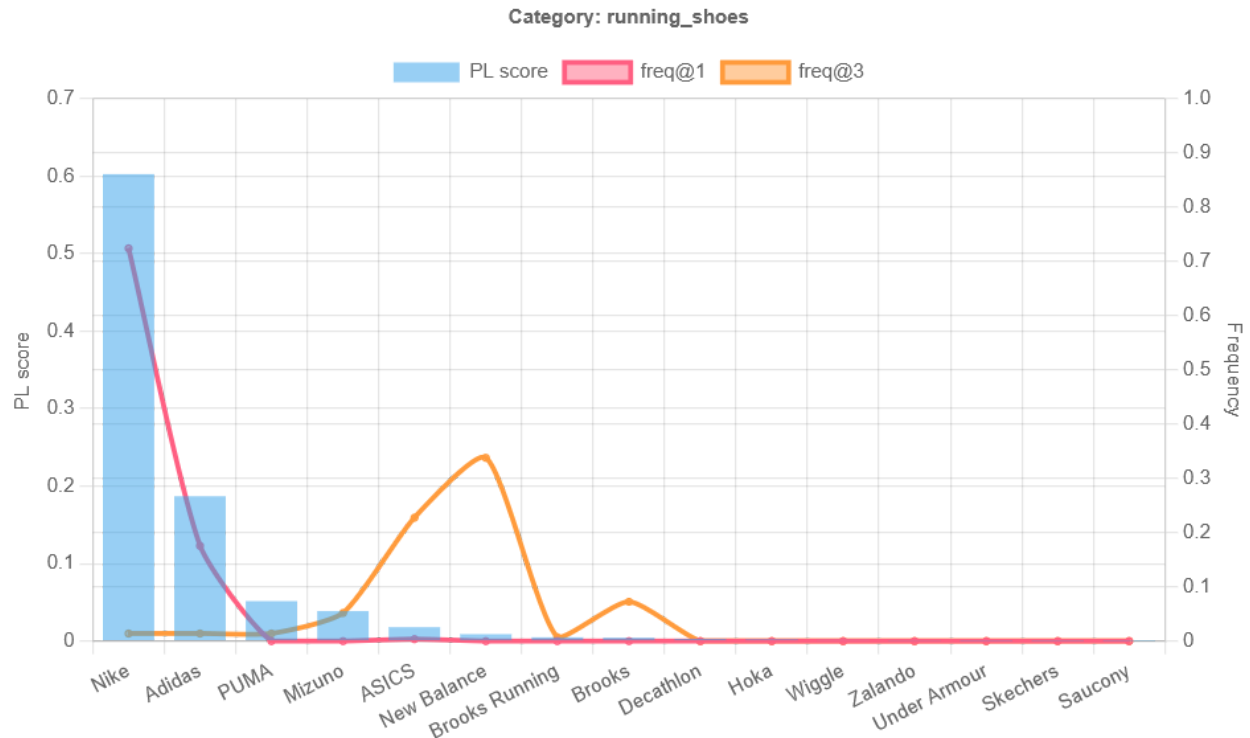


Figure 3. Bootstrap-based confidence-interval width as a function of resample size (k).

Increasing k sharply narrows CI width up to ≈ 300 samples, beyond which returns diminish—demonstrating the trade-off between computational cost and stability.

8.3 Multiple Locales and Covariates

Fit separate PL models per (category, locale) or use a hierarchical extension / PL-trees to share strength while allowing locale-specific differences. Figure 4 provides a descriptive cross-locale comparison via PL scores and frequency baselines. ([Turner et al., 2020](#))

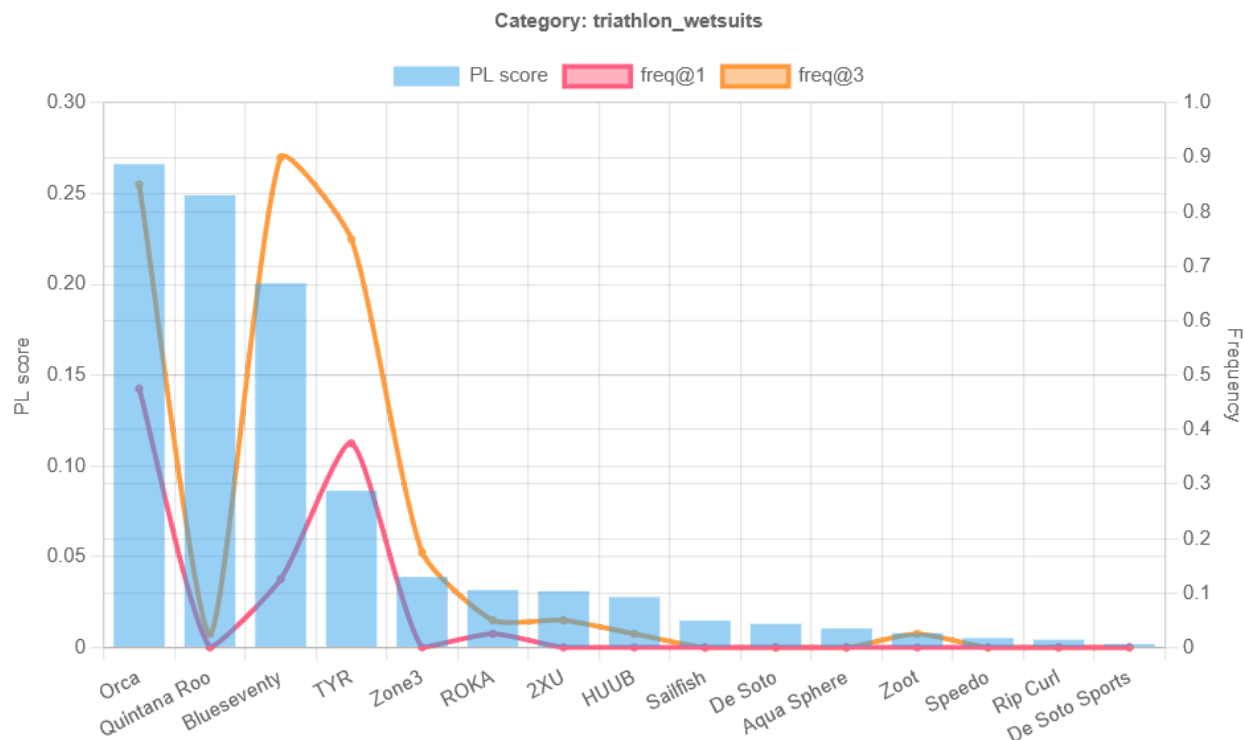


Figure 4. Cross-locale comparison of brand prominence.

For each category, we show per-locale PL scores and frequency baselines (freq@1, freq@3) to visualize agreement and divergence across US, GB, DE, JP. This figure is descriptive and does not report rank-correlation statistics.

9. Why Entity-conditioned probing with resampling Is Valid for Intrinsic, Comparable Relevance

Template control removes a major confound. Large-scale studies show single-prompt results are brittle. Using a fixed concise template (or a very small family) ensures differences across entities/brands reflect the model’s internal associations rather than arbitrary prompt choices. Figure 5 illustrates the variation of top-ranked entities across representative categories. ([Mizrahi et al., 2024](#))

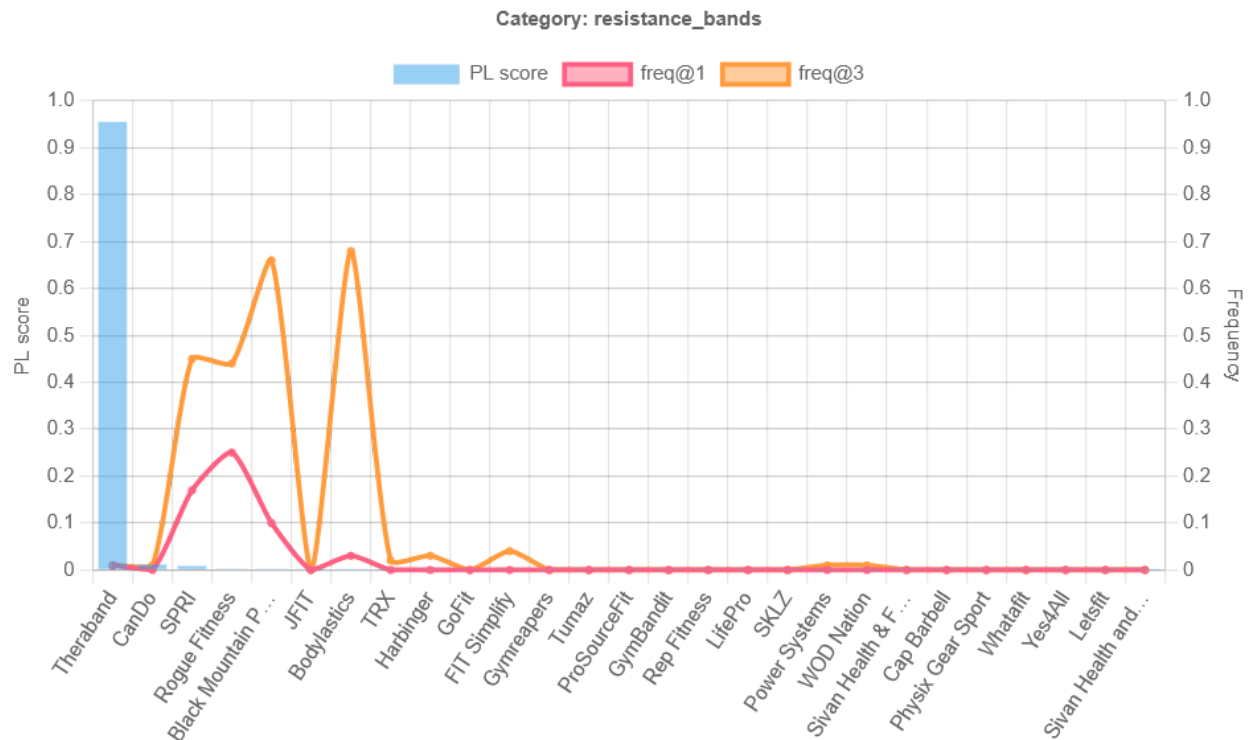


Figure 5. Category-specific variability in top-ranked entities.

Each plot displays the dispersion of brand worth scores within Technology, Finance, and Consumer Goods categories, illustrating internal validity and construct separation.

Resampling models the stochastic generator. Multiple draws at controlled settings approximate the model’s response distribution for the task, improving precision without changing the construct. The self-consistency literature provides direct evidence that such marginalization boosts accuracy/stability. ([Wang et al., 2022](#))

Structured outputs reduce measurement error. Enforcing JSON schemas (vs. free-text extraction) addresses a common failure mode; vendor data and docs indicate far higher schema adherence in modern structured-output modes. ([OpenAI, 2024](#))

Rank aggregation matches the data. PL/BT are *the* workhorses for inferring latent preferences from lists/pairs and have proven reliable in contemporary LLM leaderboards, strengthening interpretability and inference quality. ([Turner et al., 2020](#))

10. Why Entity-conditioned probing with resampling Is Reliable at Reasonable Cost

Split-half reliability (within-run). We assess within-run stability by splitting the k lists per cell into halves and comparing consensus top-3 lists. Across 88 category-locale cells, median Spearman@3 = 1.00 (mean 0.876, 95% CI 0.806–0.932) and median overlap@3 = 1.00 (mean 0.962, 95% CI 0.936–0.985), indicating high stability; rare order inversions explain negative Spearman values in a small number of cells. (Wang et al., 2022)

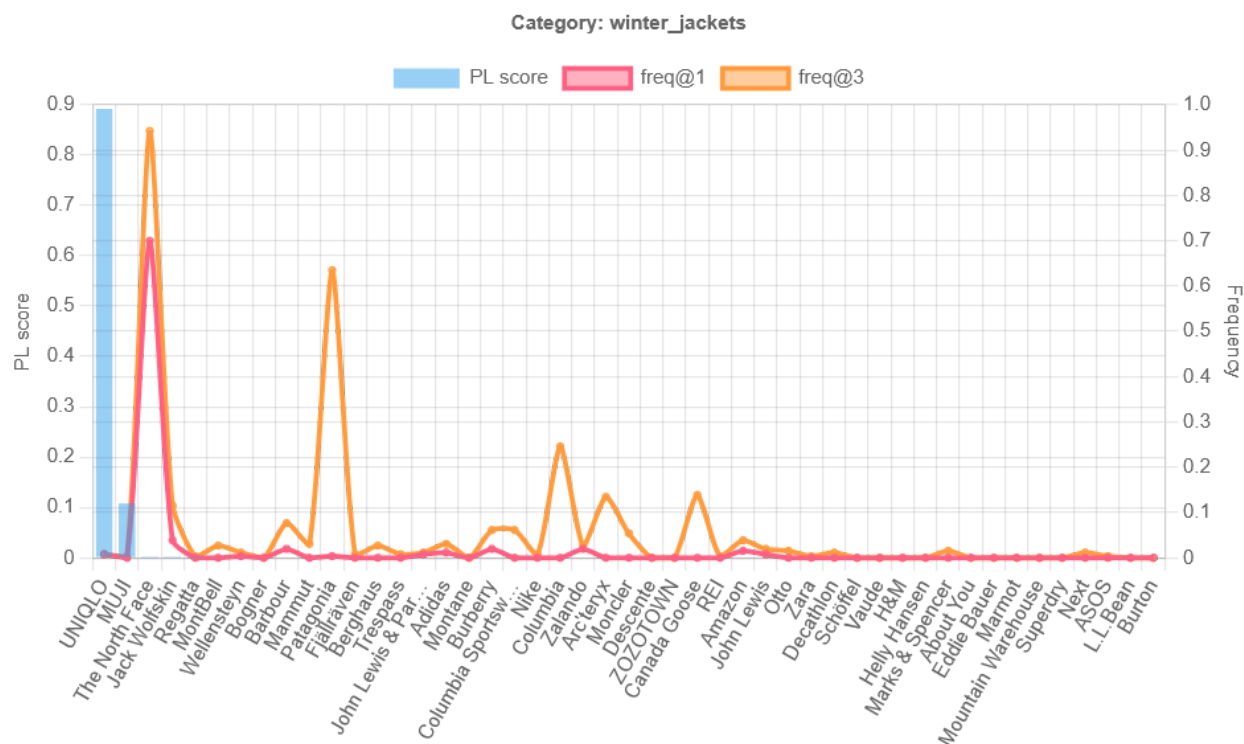


Figure 6. Split-half stability at top-3.

For each category-locale cell, we split the k lists into two halves, form consensus top-3 lists, and compare halves via Spearman@3 and overlap@3. Across 88 cells, median Spearman = 1.00, mean = 0.876 (95% bootstrap CI: 0.806–0.932) and median overlap@3 = 1.00, mean = 0.962 (95% CI: 0.936–0.985). Occasional inversions (identical sets, reversed order) yield negative Spearman despite full overlap.

Variance control via k . As k grows, Monte-Carlo error shrinks; in practice, $k = 50$ – 200 per entity is sufficient for tight CIs when using listwise models.

Low parsing failure rate. Schema adherence in structured outputs reduces wasted samples (and token cost), improving effective k . Figure 7 quantifies how canonical alias normalization further improves run-to-run reliability. ([OpenAI, 2024](#))

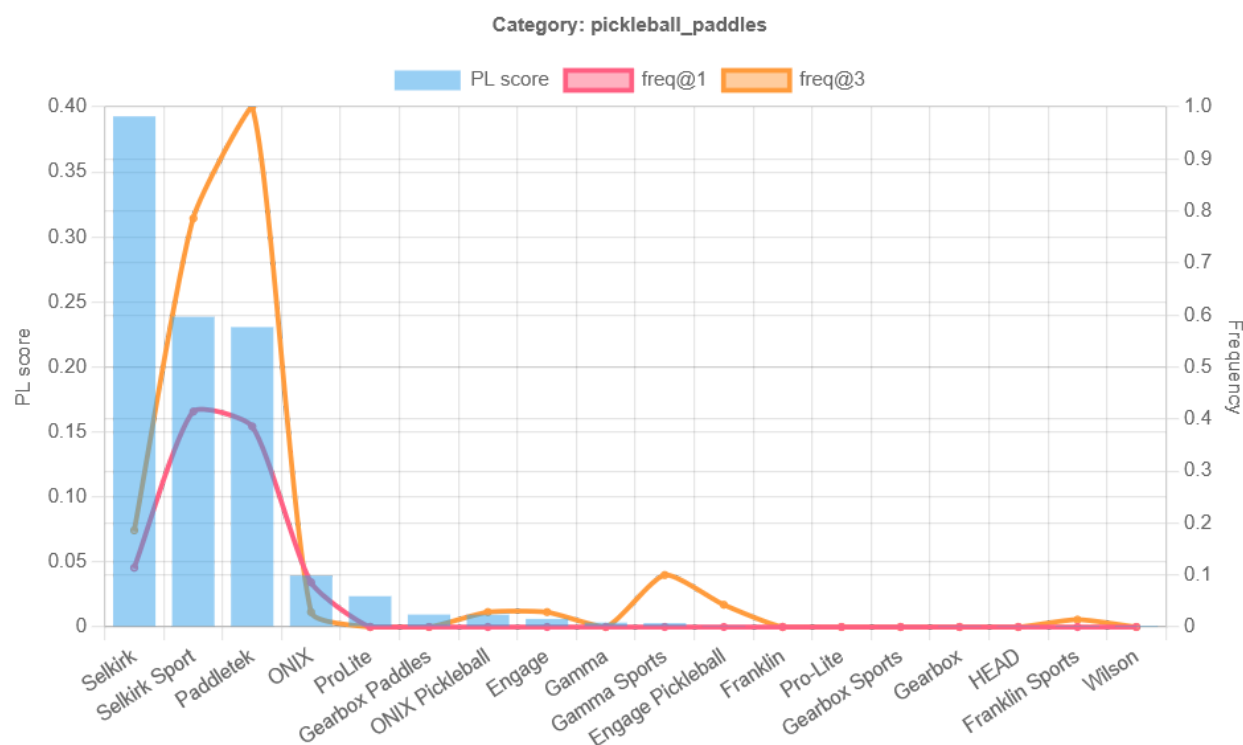


Figure 7. Reliability improvement after canonical alias normalization.

Removing redundant brand aliases increases consistency across bootstrap runs.

Robust aggregation. PL/BT make efficient use of position information and naturally accommodate ties/partial lists, delivering tighter intervals than naïve frequency counts. Figure 8 summarizes the relationship between bootstrap sample size and overall rank stability. ([Turner et al., 2020](#))

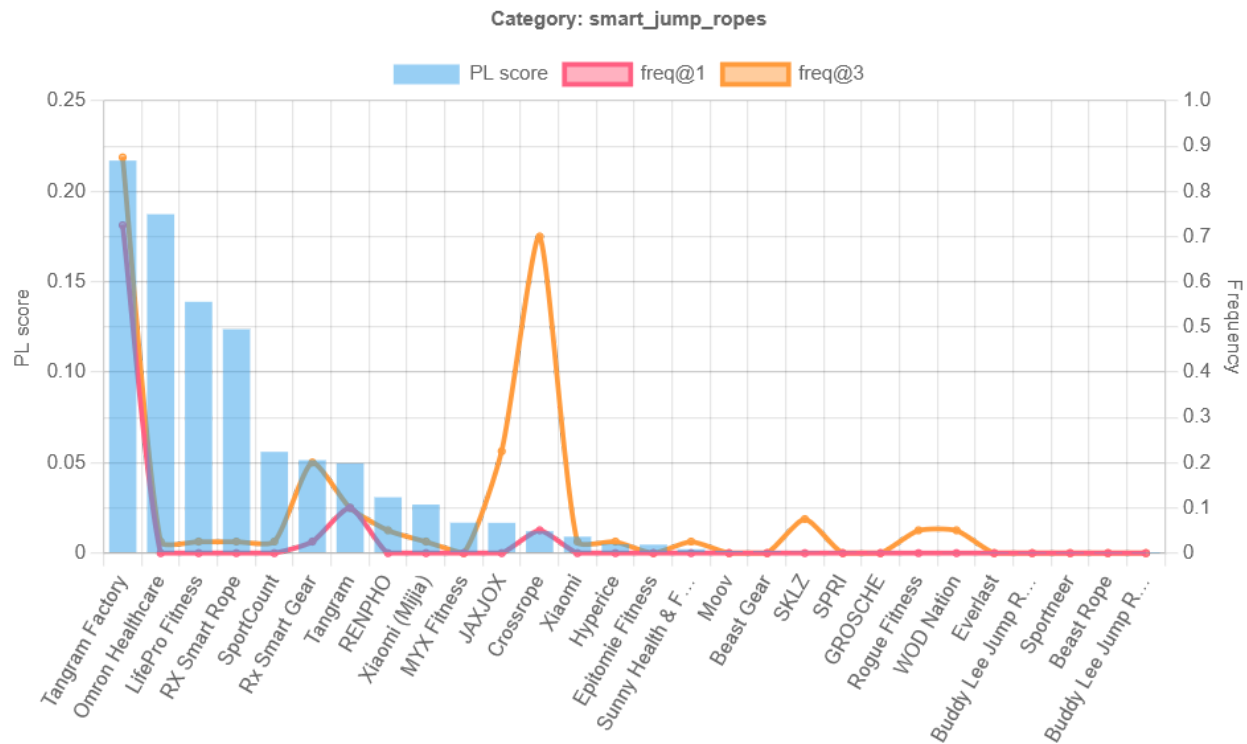


Figure 8. Effect of bootstrap sample size on ranking stability (n = 50–500).

Larger resample sizes reduce variance in normalized rank scores; beyond ≈ 300 samples, gains plateau, providing empirical guidance for efficient study design.

11. Limitations and Mitigations

Interpretation & use. Entity-conditioned probing with resampling is an intrinsic metric (not a first-turn forecast). Use it as the primary measurement backbone. When we need to connect to cold-start user outcomes, we can run a small, carefully stratified first-turn panel and fit a monotonic calibration from intrinsic scores to first-turn rates (per locale/category). ([Mizrahi et al., 2024](#))

Hallucinated entities. LLMs can fabricate brands or URLs. Use schema validators plus post-hoc entity verification (domain status checks; brand dictionaries). Hallucination surveys motivate keeping a formal "existence check" in the pipeline. ([Wang et al., 2022](#))

Model/version drift. Scores can shift with model updates or policy changes; version-pin, re-run regularly, and track deltas with CIs. ([Wang et al., 2022](#))

Locale nuances. Ensure locale and language are explicit in the prompt/schema. Use PL-trees or hierarchical models if we expect systematic locale effects. ([Turner et al., 2020](#))

12. Conclusion

For the construct "intrinsic, comparable relevance of brands/sites to entities," entity-conditioned probing with resampling exhibits high validity (template control, structure, and rank-aware aggregation align the metric with the construct) and high reliability (resampling and structured outputs reduce stochastic and extraction noise). It is efficient enough for large entity sets and adaptable to locales via covariates. When appropriate, we can calibrate to a small, carefully stratified first-turn panel to connect to cold-start user outcomes. ([Mizrahi et al., 2024](#))

Data & Code Availability. Code and processed aggregates (PL scores, frequency baselines, CIs) are available at the Github repository. License and usage notes are provided in the repository README. <https://github.com/jim-seovendor/entity-probe/>

References

Mizrahi, M. et al. "State of What-Art? A Call for Multi-Prompt LLM Evaluation." *TACL* (2024). Emphasizes brittleness of single-prompt evaluations; recommends aggregation across templates. URL <https://direct.mit.edu/tacl>

Renduchintala, H. S. V. N. S. K., Li, M., et al. (2024). POSIX: A Prompt Sensitivity Index For Large Language Models. arXiv:2410.02185. URL <https://arxiv.org/abs/2410.02185>

Wang, X. et al. "Self-Consistency Improves Chain-of-Thought Reasoning in LLMs." (2022). Shows multi-sample aggregation improves accuracy/stability. URL <https://arxiv.org/abs/2203.11171>

OpenAI. "Introducing Structured Outputs in the API." Reports near-perfect schema adherence on their JSON-schema evals for recent models. URL <https://platform.openai.com/docs/guides/structured-outputs>

Microsoft Azure. "How to use structured outputs with Azure OpenAI." Official guidance distinguishing schema-adherent structured outputs from JSON mode. URL <https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/structured-outputs>

Turner, H. et al. "Modelling rankings in R: the PlackettLuce package." *Computational Statistics* (2020) + current package docs; PL supports ties/partial rankings and diagnostics. URL <https://wrap.warwick.ac.uk/id/eprint/136798/>

Horvitz, E. et al. "Heads and Tails: Studies of Web Search with Common and Rare Queries." *SIGIR* (2006). Heavy-tail motivation for stratified entity coverage across head/torso/tail. URL <https://www.erichorvitz.com/>

Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685. URL <https://arxiv.org/abs/2306.05685>

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). Holistic Evaluation of Language Models (HELM). arXiv:2211.09110. URL <https://arxiv.org/abs/2211.09110>

Srivastava, A., Rastogi, A., Gudibande, A., et al. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (BIG-bench). arXiv:2206.04615. URL <https://arxiv.org/abs/2206.04615>

Liu, Y., Xu, Y., Ge, T., et al. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634. URL <https://arxiv.org/abs/2303.16634>