RESEARCH PAPER · NOVEMBER 2026 · LAURELIN LABS



Answering the fan-out does not predict AI Overview citation.

A pre-registered test across 235 queries: rank for the query and page length predict which pages Google's AI Overviews cite. Ranking for, or answering, the searches AI systems run behind the query does not.

PIERS BUTLER · ORCID 0009-0001-0279-9288
LAURELIN LABS LIMITED · LONDON · LAURELINLABS.COM
VERSION 1.0 · NOVEMBER 2026

ABSTRACT

AI search systems answer a question by running several related searches, a technique Google calls query fan-out. A market of tools now scores pages by how many of these sub-queries they answer, on the premise that coverage raises the chance of being cited. This study tested that premise with a pre-registered observational design. The sample was 235 non-branded keywords for which a page from one of 46 domains had been cited in a Google AI Overview. Candidate pages were the organic top ten plus every cited page. Fan-out was observed directly from Gemini with Google Search grounding (705 calls) and Claude with web search (470 calls), and guessed by a model asked to predict it (470 calls). Results pages were collected for 539 sub-queries. In conditional logistic regression within each query (158 test queries, 1,190 pages, 670 cited), a better organic rank for the head query was strongly associated with citation (odds ratio 0.43 per standard deviation of log rank) and longer pages were cited more (1.30). Heading similarity to the query, ranking in the top ten for a sub-query, and the share of sub-queries a page answers added no out-of-sample discrimination (change in within-query AUC from minus 0.003 to plus 0.004, against a pre-set bar of plus 0.02), and guessed sub-queries did as well as observed ones. Observed fan-out was small: 2.35 clusters of sub-queries per keyword against 16.1 guessed, and Gemini ran no search at all in 47% of its answers. Of cited pages outside the head query's top ten, 14% ranked in the top ten for any observed sub-query. A same-day replication on a second data provider agreed. Fan-out coverage scores are not supported as predictors of AI Overview citation; rank for the head query and length are.

KEYWORDS query fan-out, AI Overviews, answer engine optimisation, generative search, citation, information retrieval, pre-registration

Author Piers Butler, Founder, Laurelin Labs Limited, London, United Kingdom
ORCID https://orcid.org/0009-0001-0279-9288
Contact piers@laurelinlabs.com · laurelinlabs.com
Version 1.0, November 2026 · pre-registration dated 2 November 2026
Licence Creative Commons Attribution 4.0 International (CC BY 4.0)
Cite as Butler, P. (2026). Answering the fan-out does not predict AI Overview citation: a pre-registered test across 235 queries. Laurelin Labs, London.

1. Introduction

Google's guidance for site owners states that AI Overviews and AI Mode may use a "query fan-out" technique, issuing multiple related searches across subtopics and data sources to develop a response [1]. The idea has travelled fast. Within a year a set of free and paid tools appeared that take a keyword, list the sub-queries an AI system might run, check a page for the ones it does not answer, and in some cases compare two pages on how many each covers. The working assumption behind all of them is that a page answering more of the fan-out is more likely to be cited.

The published evidence for that assumption is mixed. Surfer pulled the top ten pages for 10,000 keywords, generated 33,000 fan-out queries with Gemini, and reported that pages ranking for the main query and at least one fan-out query accounted for 51.2% of AI Overview citations that rank organically, against 19.6% for pages ranking for the main query alone; it also found that 67.8% of citations ranked in the top ten for neither [2]. AirOps and Kevin Indig captured ChatGPT's own sub-queries across 16,851 queries and three runs each and found retrieval rank to be the dominant signal, with heading similarity to the query mattering and pages covering 26 to 50% of sub-queries cited more often than pages covering all of them [3]; a

companion AirOps analysis reported that 32.9% of cited pages appeared only in results for a fan-out query [4]. Ahrefs found that the share of AI Overview citations coming from the organic top ten fell from 76% in July 2025 to 38% in March 2026 and linked part of the change to fan-out [5].

Three gaps remain. First, the study of Google AI Overviews used model-generated sub-queries rather than the searches the system actually ran, so the gap between guessed and observed fan-out is unmeasured. Second, most reported figures are shares of citations, which compare cited pages with each other rather than cited pages with the uncited pages they competed against for the same query. Third, none of the studies was pre-registered, so the analysis choices could have followed the data.

This paper reports a pre-registered test of four claims that fan-out tools rely on, on Google AI Overviews, with fan-out observed from two systems and guessed by a third, and with every comparison made within the query. It also reports how large observed fan-out is, how stable it is from run to run, and how much of it a model can guess. The work was carried out to decide what Laurelin Labs' own audit engine and client platform may claim; the decision rules for that were written before the data was collected.

2. Methods

2.1 Pre-registration

Hypotheses, the sample rule, the features, the models, the pass criteria and the decision rules were written in an engine specification dated 2 November 2026, before any outcome data was pulled [6]. Changes after that point were logged with the date and reason. Eleven entries were made. The substantive ones are reported in section 4; none altered a hypothesis, a pass criterion or a decision rule.

The hypotheses were:

- **H1, rank.** Among the candidates for a query, a better organic rank for the head query raises the odds of citation. This was a check on the data.
- **H2, retrievable for sub-queries.** A page in the top ten for at least one observed sub-query is more likely to be cited, over and above its head rank.
- **H3, heading match.** The similarity of a page's best heading to the head query raises the odds of citation, over and above rank.
- **H4, coverage.** The share of observed sub-queries a page's text answers adds to the prediction from rank, heading match and H2. A secondary test asked whether pages covering 26 to 50% are cited at least as often as pages covering 76 to 100%.
- **H5, observed against guessed.** Coverage built from observed sub-queries predicts citation better than coverage built from guessed ones.

2.2 Sample

The frame was a corpus of 1,795 keyword rows for which Ahrefs reported a page from one of the client and competitor domains Laurelin Labs has audited as cited in a Google AI Overview (United Kingdom and United States, pulled 11 October 2026). Rows flagged branded or navigational were removed, leaving 1,252 unique keyword and country pairs. At most six were drawn per cited domain with a fixed random seed, giving 235 pairs across 46 domains (162 United Kingdom, 73 United States). The first 40 in random order formed a calibration fold used only to set a threshold; the other 195 formed the test fold.

2.3 Citation labels and candidate pages

For each keyword the latest Ahrefs results page was pulled for its country: organic positions with domain rating, AI Overview cited URLs, and People Also Ask questions. A candidate page was any organic result in the top ten or any cited URL, and a candidate was labelled cited if it appeared as an AI Overview citation. 207 of the 235 keywords had an AI Overview in the snapshot. The 235 pulls were made by eight parallel software agents; six keywords chosen at random were pulled again directly and matched on every URL and position.

2.4 Fan-out capture

Observed fan-out came from two systems that report the searches they run. Gemini (model reported as gemini-3.6-flash) was called with the Google Search tool three times per keyword, and the searches were read from the grounding metadata. Claude Haiku 5.5 was called with the web search tool twice per keyword, and the searches were read from the tool calls. In test calls Claude often answered without searching, so it was made to search at least once; how many further searches it ran was left to the model. Gemini decided for itself. Each call carried one system line naming the user's country. Guessed fan-out came from Claude Haiku 5.5 without tools, asked twice per keyword to list the 8 to 12 searches Google's AI Mode would run, which is the method most free tools use.

2.5 Normalising and clustering sub-queries

Sub-queries were lower-cased and normalised, with the current and next year removed. Within each keyword, sub-queries whose embeddings were at cosine 0.90 or more were merged into one cluster (bge-small-en-v1.5 [7]). A cluster's stability for a source is the share of that source's runs in which it appears; overall stability is the highest of the per-source values.

2.6 Results pages for sub-queries

For each keyword, up to ten observed clusters (most stable first) were searched on Google through DataForSEO's standard queue, depth ten, in the keyword's country, on 3 November 2026: 540 tasks, 539 returned. 2,945 of the 4,624 organic result URLs came back wrapped as google.com/goto links with opaque tokens. Each was resolved with one request to www.google.com, which answered with a redirect to the page; all 2,945 resolved.

2.7 Page text and features

Candidate pages were fetched once and read with the extractor of Laurelin Labs' audit engine: consent dialogues and navigation removed, main content located, headings H1 to H4 kept, and passages of 220 words with 40 words of overlap built at heading breaks. While building the study the extractor was found to discard all content on pages made with the Elementor page builder, because its wrapper class names contain the word widget; this was fixed before any feature was computed. 1,553 of 2,370 candidate URLs returned text. Fetch success was 66% for cited and 64% for uncited candidates.

Queries were embedded with the model's query instruction and passages and headings without it. For each page and keyword the features were: log of organic rank (pages outside the top ten set to log 11, with an indicator), log word count, domain rating, the best heading's cosine to the head query, whether the page was in the top ten for any observed sub-query and the share of sub-queries for which it was, and coverage, the share of sub-query clusters whose best passage reached a cosine threshold tau. Coverage was computed separately for observed and for guessed sub-queries.

tau was fixed on the calibration fold. The plan was to set it where a model judge's "this passage answers this search" reached 0.8 precision on 150 labelled pairs. The judge marked only 15 of 144 labelled pairs as a clear yes, against 93 partly and 36 no, so precision never reached 0.8 and the pre-registered fallback applied: tau equals 0.798, the 75th percentile of best-passage cosines in the fold.

2.8 Models and tests

The unit was a candidate page within a keyword, and only keywords with at least one cited and one uncited candidate entered. The primary method was conditional logistic regression grouped by keyword [8], so that keyword difficulty cancels out. Models were nested: M0 with rank, word count and domain rating; M1 adding heading match; M2 adding the sub-query rank features; M3 adding observed coverage; and M3p adding guessed coverage in its place. Out-of-sample discrimination was the within-keyword AUC from five-fold cross-validation grouped by keyword, with 95% intervals from 2,000 bootstrap resamples of keywords. H2 and H4 required a positive coefficient and an AUC gain whose interval lay above zero, and for H4 a gain of at least 0.02. Holm correction was applied across H2, H4 and H5.

The decision rules were set in advance. If H4 passed, a coverage score could be shown as related to citation. If H4 failed and H2 passed, products could lead with findability for sub-queries but show coverage only as a checklist. If both failed, sub-queries could be used as content research only, with no score. Heading advice was supported only if H3 passed.

2.9 Independent review

Before the result was used, a separate AI agent with no access to the author's interpretation reviewed the code, data and results and re-computed the headline figures. It found the URL wrapping described in 2.6 (at that point unresolved), a lookup fault that left People Also Ask coverage mostly empty, and the selection artefact described in 3.3. The wrapping and the lookup fault were fixed and every analysis re-run; the results below are the corrected ones. The decision was the same before and after.

3. Results

3.1 Sample flow

Stage	Keywords	Pages
Eligible keyword and country pairs	1,252	
Drawn (at most six per cited domain, 46 domains)	235	
With an AI Overview in the Ahrefs snapshot	207	
Candidate URLs (organic top ten plus cited)		2,370
Candidate URLs with text		1,553
Test fold, at least one cited and one uncited candidate, complete features	158	1,190 (670 cited)

Table 1. Sample flow.

3.2 What the systems searched for

Measure	Value
Gemini: searches per answer	0.88
Gemini: answers with no search at all	331 of 705 (47%)
Gemini: keywords with no search in any of three runs	77 of 235
Claude (made to search at least once): searches per answer	1.08
Claude: answers with exactly one search	433 of 470
Observed sub-query clusters per keyword, both systems	2.35
Guessed sub-queries per keyword (after merging)	16.1
Gemini clusters seen in all three of its runs	16.5%
Claude clusters seen in both of its runs	79.8%
Clusters seen by both systems	115 of 542 (21%)
Observed clusters with a near match in the guessed list (cosine 0.90 or more)	58.4%

Table 2. Fan-out as observed and as guessed. Figures per keyword are over the 170 keywords in the analysis frame.

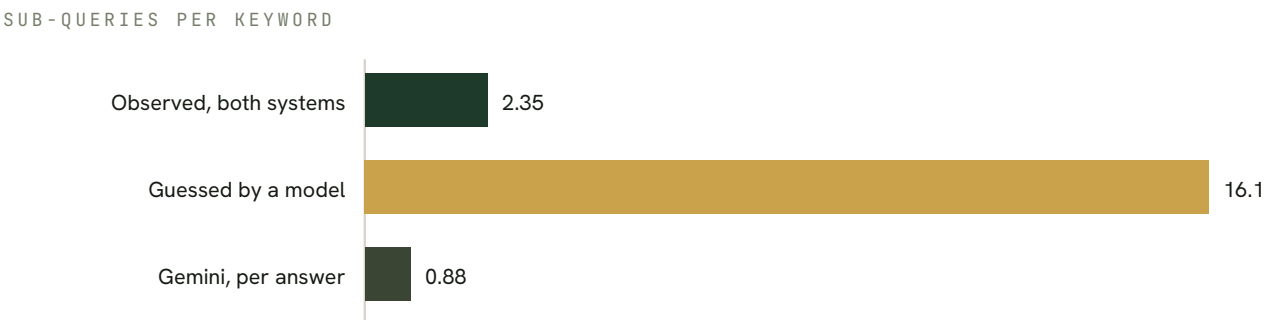


Figure 1. The searches AI systems actually ran against the number a model guesses. A guessed list is about seven times longer than the observed set of distinct searches.

3.3 Primary models

Every candidate outside the organic top ten is in the set only because it was cited, so the indicator for being outside the top ten separates the outcome perfectly. Those pages contribute no information about the other features, and the coefficients below are in effect estimated on pages inside the top ten. A refit on those 662 pages alone (122 keywords) gave the same coefficients.

Feature	M1	M2	M3	M3p
Log organic rank (per SD)	0.44 *	0.44 *	0.43 *	0.43 *
Log word count (per SD)	1.34 (p 0.006)	1.34 (p 0.006)	1.30 (p 0.015)	1.24 (p 0.048)
Domain rating (per SD)	0.92 (p 0.49)	0.92 (p 0.50)	0.94 (p 0.59)	0.94 (p 0.59)
H3: best heading match to the query (per SD)	0.87 (p 0.26)	0.87 (p 0.25)	0.81 (p 0.11)	0.79 (p 0.06)
H2: top ten for an observed sub-query		1.37 (p 0.59)	1.30 (p 0.65)	1.36 (p 0.60)
H2: share of sub-queries in the top ten (per SD)		0.94 (p 0.78)	0.97 (p 0.89)	0.96 (p 0.87)
H4: observed coverage (per SD)			1.29 (p 0.029)	
H5: guessed coverage (per SD)				1.52 (p 0.002)

Table 3. Odds ratios from conditional logistic regression within keyword (158 keywords, 1,190 pages). * $p < 0.0001$. An odds ratio below 1 for log rank means a lower position number (a better rank) goes with more citations.

3.4 Out-of-sample discrimination

Comparison	AUC before	AUC after	Change	95% interval
Add heading match (M1 against M0)	0.874	0.878	+0.004	−0.000 to +0.009
Add sub-query rank (M2 against M1)	0.878	0.878	−0.001	−0.005 to +0.004
Add observed coverage (M3 against M2)	0.878	0.875	−0.003	−0.009 to +0.003
Observed against guessed coverage (M3 against M3p)	0.873	0.875	+0.001	−0.009 to +0.012
Pages inside the top ten only (662 pages, 122 keywords)				
Add heading match	0.741	0.744	+0.003	−0.009 to +0.015
Add sub-query rank	0.744	0.745	+0.001	−0.008 to +0.010
Add observed coverage	0.745	0.739	−0.006	−0.025 to +0.012

Table 4. Within-keyword AUC from five-fold cross-validation grouped by keyword, with bootstrap intervals over keywords. The bar for H4 was a gain of at least +0.02 with an interval above zero. The base AUC of the full set (0.874) is raised by the perfect separation described in 3.3; the top-ten-only figures are the fairer reading.

H1 passed. H2, H3, H4 and H5 failed. Observed coverage had a positive coefficient within the sample (odds ratio 1.29 per standard deviation, p 0.029), but it carried no information that rank and length did not already carry, so it did not improve prediction on held-out keywords. Guessed coverage behaved the same way and was, if anything, the stronger in-sample correlate.

3.5 Coverage bands

SHARE OF PAGES CITED, BY SHARE OF OBSERVED SUB-QUERIES ANSWERED

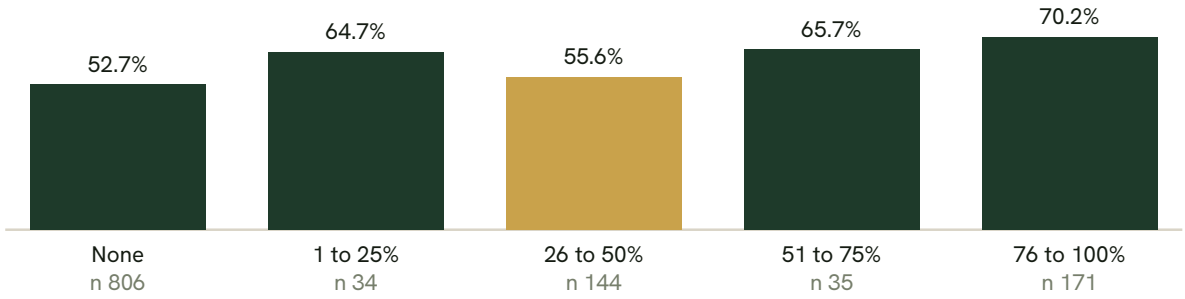


Figure 2. Pages answering most of the observed sub-queries were cited most often, the reverse of the "focused beats exhaustive" finding for ChatGPT [3]. Within keyword, the top band's coefficient against the 26 to 50% reference was +0.33 (p 0.025). The pattern does not survive as prediction once rank and length are known (Table 4).

3.6 Robustness

Variant	Keywords	Add sub-query rank	Add coverage	Decision
Primary	158	−0.001	−0.003	No score
People Also Ask added to observed	158	−0.001	−0.007	No score
Gemini sub-queries only	100	−0.004	−0.005	No score
tau 0.748	158	−0.001	+0.003	No score (H5 passes)
tau 0.848	158	−0.001	−0.004	No score
Platforms removed (video, social, forums)	155	−0.001	+0.002	No score
United Kingdom only	107	−0.003	−0.004	No score

Table 5. Changes in within-keyword AUC under each variant. A Claude-only model could not be fitted, because most Claude runs made a single search, usually close to the head query.

3.7 Exploratory analyses (not pre-registered)

Citations outside the top ten. 54% of cited pages lay outside the head query's organic top ten. Of these, 14% ranked in the top ten for an observed sub-query, compared with 20% of cited and 16% of uncited pages inside the top ten. Observed sub-queries do not explain where the citations from outside the top ten come from.

Same-day replication. The time between the Ahrefs snapshots (October) and the sub-query results (3 November) is a weakness: for the 74 keywords that were also their own sub-query, the two providers' top ten shared 13.5% of URLs and 30% of domains. The head results were therefore pulled again from DataForSEO on 3 November with AI Overview references, giving labels and sub-query results from one provider on one day. 190 keywords had an AI Overview. Among top-ten pages (845 pages, 102 keywords) rank replicated (odds ratio 0.60 per standard deviation, $p < 0.0001$); ranking in the top ten for a sub-query had an odds ratio of 1.60 (p 0.051) and changed within-keyword AUC by −0.012 (−0.027 to +0.002). Before the organic URLs in these results were resolved, the same model showed an odds ratio of 7.8, an artefact of the wrapped links.

4. Discussion

The test was built to give fan-out coverage a fair chance. Sub-queries were observed rather than guessed, the comparison was between pages competing for the same query, the threshold for "answers" was set on held-out data, and the pass bar for coverage was modest. Coverage failed it, and so did ranking for sub-queries and heading match. What predicted which top-ten pages AI Overviews cite was their rank for the head query and, to a lesser degree, their length.

The result does not contradict the Surfer figures so much as answer a different question. Shares of citations describe where cited pages come from; they do not show that a page answering more sub-queries beats the pages it competed with. When cited pages are compared with uncited pages for the same query, the sub-query features add nothing. The AirOps findings for ChatGPT, that heading match helps and exhaustive coverage hurts, did not replicate for AI Overviews: heading match was flat, and the most complete pages were cited most often, while still adding no predictive value beyond rank and length. Retrieval rank, the strongest signal in that study, was the strongest signal here too.

The size of observed fan-out matters for practice. The searches the two systems ran averaged 2.35 distinct clusters per keyword, close to the two sub-queries AirOps observed for most ChatGPT queries [3], while a model asked to guess produced 16. Gemini ran no search at all in nearly half its answers, and when it did search, only one cluster in six recurred in all three of its runs. A tool that shows a long list of guessed sub-queries is presenting about seven times more searches than the systems ran, most of them unstable, and a page edited to answer all of them is being edited against a list that does not exist in that form. Some guessed sub-queries do match observed ones (58% of observed clusters had a near match), so the guessed list is not noise; it is a superset.

For practitioners this supports three things. The searches around a query remain useful as content research, because they describe what people and systems look for. Ranking for the head query remains the main route into the pool of pages AI Overviews cite. And scores that claim to predict citation from sub-query coverage should be treated as unvalidated until shown otherwise on comparisons within the query.

Two measurement lessons are worth recording. The candidate set used here, top ten plus cited pages, makes every page outside the top ten cited by construction, which inflates any model's overall AUC; a fair reading needs the top-ten-only refit or a candidate set that includes uncited pages from deeper positions. And a widely used provider of results-page data was, at the time of the study, returning most result URLs as wrapped links, which silently breaks any analysis that matches pages by URL or domain. In this study it first hid real matches and then, in the same-day replication, manufactured a strong false effect.

5. Limitations

- The design is observational. A pass would have shown association, not cause; a fail does not prove that adding coverage never helps a particular page.
- Gemini's API searches are a stand-in for the searches AI Overviews run internally. Claude's searches are not Google's at all.
- Ahrefs samples AI Overviews, and its snapshots predate the fan-out capture by days to weeks. The same-day replication on a second provider addresses this for H1 and H2 only.
- The keywords are ones for which a domain Laurelin Labs has audited was cited, so the sample leans towards the sectors of those clients and their competitors, and towards queries where an AI Overview appears.

- Many keywords have only one or two observed clusters, so coverage takes few values and the test of H4 has limited power for small effects.
- A third of candidate pages could not be read, mostly because they blocked the fetcher or render their content with scripts; the failure rate was similar for cited and uncited pages.
- Results pages and AI Overviews change from day to day. The figures describe October and November 2026.

6. Conclusion

> On 158 queries, answering or ranking for the searches AI systems run behind a query did not predict which pages Google's AI Overviews cite, once the page's rank for the query and its length were known. Real fan-out was small and unstable, and a guessed list was about seven times longer than what the systems searched for.

Under the decision rules set before the study, Laurelin Labs' audit engine and client platform present sub-queries as content research, without a coverage score or a findability score for citation, and treat heading advice as a writing convention. The study will be repeated when the engine's own observed fan-out runs have accumulated across client sites, with a candidate set that includes uncited pages from deeper positions.

Data and code availability

The specification with its deviations log, the analysis code, the sample, the analysis frame (one row per keyword and candidate page, with URLs and features but no page text), and the result files are deposited with this paper. Page text and raw provider responses are not redistributed. The audit engine code that reads pages and runs the fan-out module is proprietary; the parts used here (text extraction, passage building and URL resolution) are described in section 2.

Competing interests

The author founded and directs Laurelin Labs Limited, which sells search and AI visibility consultancy and builds the audit engine and platform whose product claims this study was designed to decide. The decision rules were fixed in advance so that the commercial interest could not choose the outcome. No external funding was received. Data and model costs were met by Laurelin Labs: about 55,000 Ahrefs units, about USD 25 of Gemini calls, about USD 17 of Claude calls, about USD 4 of OpenAI calls for an arm that was not completed and is not reported, and under USD 1 of DataForSEO tasks.

Use of AI tools

The study was designed, coded and run with Claude (Anthropic) working under the author's direction in the author's own environment, and the independent review in section 2.9 was carried out by a separate Claude agent. The author set the question, approved the design and is responsible for the interpretation and for any errors.

References

1. Google Search Central. AI features and your website. https://developers.google.com/search/docs/appearance/ai-features (accessed 3 November 2026).
2. Hardwick, J. (2025, 28 November). Ranking for multiple fan-out queries dramatically increases your chances of getting cited in AIOs (173,902 URLs studied). Surfer. https://surferseo.com/blog/query-fan-out-impact
3. AirOps and Indig, K. (2026, April). The fan-out effect: what happens between a query and a citation. AirOps. https://www.airops.com/lp/the-fan-out-effect-report
4. AirOps (2026, 12 March). The influence of retrieval, fan-out, and Google SERPs on ChatGPT citations. https://www.airops.com/blog/query-fanout-pipeline
5. Linehan, L. (2026, 2 March). Update: 38% of AI Overview citations pull from the top 10. Ahrefs. https://ahrefs.com/blog/ai-overview-citations-top-10
6. Butler, P. (2026). SPEC-1.45.0: query fan-out, pre-registered study, engine module and free tool, with deviations log. Laurelin Labs. Deposited with this paper.
7. Xiao, S., Liu, Z., Zhang, P. and Muennighoff, N. (2023). C-Pack: packaged resources to advance general Chinese embedding. arXiv:2309.07597. Model BAAI/bge-small-en-v1.5.
8. Seabold, S. and Perktold, J. (2010). Statsmodels: econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference. ConditionalLogit, statsmodels.discrete.conditional_models.