

HOW MANY QUERIES DO YOU NEED? BUDGET-GOVERNED LABEL SPARSITY IN SERP-DERIVED WEB SUPERVISION

Muhammad Mohtasham Mujahid^{*1}, Fawad Nasim²

^{*1}Faculty of Computer Science and information Technology, The Superior University, Lahore, Pakistan

^{*1}mohtashamdar1998@gmail.com, ²fawad.nasim@superior.edu.pk

DOI: <https://doi.org/10.5281/zenodo.22126098>

Keywords

search engine result pages, weak supervision, label sparsity, sampling budget, web measurement, class imbalance, reproducibility

Article History

Received: 28 April 2026

Accepted: 15 June 2026

Published: 30 June 2026

Copyright ©Author

Corresponding Author: *
Muhammad Mohtasham
Mujahid

Abstract

Work that predicts which web domains surface in search results tends to build its labels the same way: issue a set of queries, record what comes back, treat everything unseen as negative. How many queries gets reported as a detail, and is often settled by whatever a vendor's free tier permits. It is not a detail. Holding a 1,000-domain frame and its features fixed while sweeping the budget from 5 to 500 queries, we find labelled positives growing as $K^{0.47}$ against $K^{0.87}$ for distinct domains seen. A popularity frame exhausts; the open web does not. Labels therefore cost about the square root of spend. Budget also decides what a study concludes, not merely how tightly. A fused model over Chrome UX Report, Common Crawl and live-crawl features loses to a zero-parameter popularity ranker at small budgets, draws level near 25 queries, and wins by 0.149 PRAUC at 500. Same code, same features, opposite claim. Leave-one-source-out ablation puts the entire effect in Common Crawl together with CrUX; the live crawl, by a wide margin the costliest source we collected, adds nothing we can measure. Two labelling decisions usually compressed into a single clause, domain resolution and rank cutoff, shift the measured effect by nearly a factor of two. Re-issuing those same queries four months on returns result sets overlapping at mean Jaccard 0.49, so supervision depreciates too. We close with a formula converting a target prevalence into a query budget, and with two budgets below which claims about feature value outrun the data that produced them. Code, data and environment are released.

I. INTRODUCTION

A recurring pattern in applied search analytics is to treat search engine result pages as ground truth. A set of queries is issued, the returned domains are recorded, and a model is trained to predict which domains appear. The methodological description usually occupies one sentence, and the number of queries is reported as an incidental detail, often fixed by whatever an API vendor's free tier happened to allow.

That number is not incidental. A budget of K queries returning ten results each observes at most 10K domain slots, and those slots are spread thin over a heavily skewed distribution.

Every domain that does not appear is recorded as a negative, whether it genuinely never ranks or simply was not asked about. The resulting label set is not a noisy view of the truth; it is a systematically incomplete one, and its incompleteness is controlled by a parameter that practitioners set almost arbitrarily.

We study that parameter directly. Holding a fixed 1,000domain frame and a fixed set of public domain features constant, we vary only the query budget used to construct labels and measure what moves. This isolates the sampling instrument from everything else in the pipeline.

Objectives

- 1) We quantify how the SERP query budget governs the prevalence, stability, and learnability of SERP-derived supervision for web-domain modeling.
- 2) We test, by leave-one-source-out ablation across four public sources, whether multi-source fusion measurably improves organic SERP-presence prediction over any single source.
- 3) We derive and empirically validate the query budget required for stable supervision, giving practitioners a concrete sampling target instead of an arbitrary one.

Contributions

Supervision yield obeys a square-root law. Labelled positives grow as $6.21K^{0.47}$ ($R^2 = 0.993$) across a hundredfold range of budgets, placing the exponent squarely in the 0.4–0.6 band Heaps' law exhibits for vocabulary growth [1]. Doubling the query budget buys $1.4\times$ the labels, not twice as many. Distinct domains observed, by contrast, grow at $K^{0.87}$: the open web does not saturate, but a fixed popularity frame does. Inverting the fit gives a closed-form budget for a target prevalence.

Conclusions about feature value invert with budget. At $K = 10$ a fused model sits 0.039 PR-AUC behind a popularity ranker that fits no parameters at all. By $K = 500$ it is 0.149 ahead. The crossing happens near $K \approx 25$. Nothing about the code or the features changed between those two readings. Which signals appear to matter is equally unstable: at the budget a free API tier affords, the importance ranking agrees with the full-budget one at only Kendall $\tau = 0.62$. Fusion helps, but two of four sources do all the work. Given enough labels the fused representation does beat popularity, 0.329 against 0.180 PR-AUC at permutation $p = 0.0099$. Drop Tranco rank from it and the score rises. Drop the live crawl, much the most expensive source to collect, and nothing moves. What is left is Common Crawl and CrUX, complementing each other.

Supervision decays. Re-issuing the same queries four months later returns result sets with mean Jaccard 0.49; the induced positive-label set has

Jaccard 0.69. A query budget is a depreciating asset, not a one-time purchase.

All code, collected data, and the environment lockfile needed to reproduce these results are released with the paper.

II. RELATED WORK

Sampling frames for web measurement. Popularity lists underpin a large fraction of empirical web research, and their properties propagate into every result computed over them. Le Pochat et al. showed that commercial rankings can be manipulated cheaply and introduced Tranco as a reproducible alternative [2]. Scheitle et al. measured the day-to-day instability of such lists and found that results computed over them systematically overestimate relative to the general web population [3]. Both lines of work interrogate the frame. We take the frame as given and interrogate the labels attached to it, which has received far less attention despite being equally consequential.

Supervision from search results. Treating retrieved result pages as ground truth is common in applied search analytics and is usually described in a sentence. What that sentence conceals is that the query sample is a sampling instrument with its own statistics. Weak-supervision work models label noise [4]; the failure mode here is label absence under a finite observation budget, which behaves differently. An unobserved positive is indistinguishable from a true negative, and no amount of modelling recovers it.

Evaluation under class imbalance. Davis and Goadrich showed that dominance in ROC space does not imply dominance in precision-recall space [5], and Saito and Rehmsmeier that ROC curves mislead at low prevalence [6]. We supply a concrete instance: a single-source model reaching ROC-AUC 0.83 while ranking below a zero-parameter popularity baseline on PR-AUC. Scaling laws for discovery and learning. Heaps' law describes vocabulary growth in text corpora as $V = aN^b$, with b usually landing between 0.4 and 0.6; Egghe worked through the formulation and its relatives in detail [1]. Ecology asks a neighbouring question, how many classes were never sampled at all, and answers it with richness estimators [7]. A third literature fits power laws to learning curves [8]. These have not been read

together. They should be: domain discovery under a query budget turns out to obey a Heaps-type law, and it is the label yield that law produces, rather than the row count, that bounds what any model on the frame can achieve.

III. DATA AND METHOD

A. Sampling Frame and Feature Sources

The unit of analysis is a registrable domain. We take the Tranco top-1,000 as the frame [2], chosen because it is reproducible, versioned, and explicitly hardened against the manipulation

that affects commercial popularity lists. Its known biases toward infrastructure and platform domains are a limitation we return to in Section V.

Figure 1 summarises the design. Four families of public signals are attached to each domain, none requiring site ownership or privileged access:

- TRANCO: the popularity rank itself, one feature.
- CRUX: field Core Web Vitals from the Chrome UX

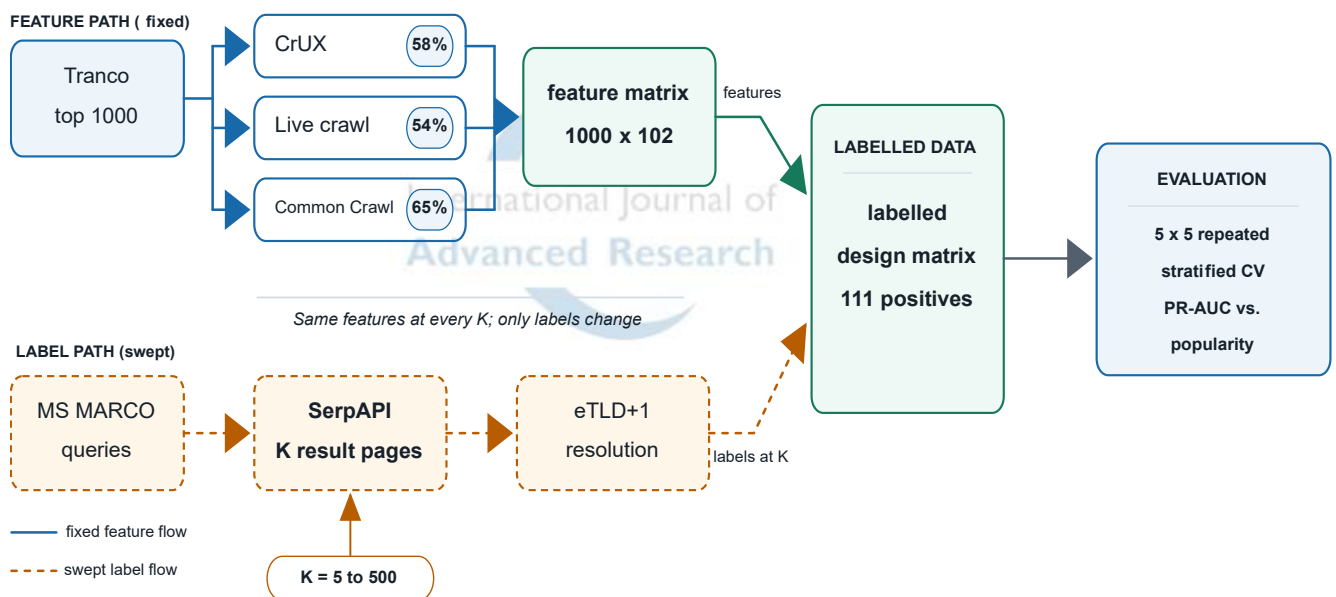


Fig. 1. Study design. The feature path is computed once and held fixed; only the label path depends on the query budget K . Percentages are the share of the 1,000-domain frame each source covers. Solid and dashed borders distinguish the two paths independently of colour.

Report [9]: p75 values and good/needs-improvement/poor shares for LCP, FCP, INP, CLS and TTFB, split by form factor. We collect INP rather than FID, the latter having been withdrawn from CrUX in September 2024 [10].

- CRAWL: a polite live crawl of up to five pages per domain, aggregated to domain level: HTTPS share, viewport and hreflang presence, JSON-LD counts, title and metadescription lengths, heading counts, internal and external link counts, word counts, and canonical adoption.
- CC: Common Crawl [11] CDX index aggregates, covering record and unique-URL counts, HTTP status-code shares, MIME-type shares, and content-length statistics.

This yields 102 numerical features per domain: 61 from CrUX, 28 from the live crawl, 12 from Common Crawl, and the Tranco rank.

Coverage is uneven and the unevenness is itself informative.

CrUX returns a record for 58.0% of the frame, Common Crawl for 65.0%, and 53.6% of domains served at least one crawlable page despite 99.5% being reachable for a crawl attempt. A domain absent from CrUX is not missing at random: it is a domain with too little Chrome traffic to clear Google's anonymity threshold.

B. Label Construction

Supervision comes from Google result pages retrieved through SerpAPI [12] for a random sample of MS MARCO queries [13], drawn with a fixed seed. Each result page yields up to ten organic results; a frame domain is labelled positive if it appears in at least one page.

Two choices in that sentence turn out to carry more weight than their length suggests, so we treat both as experimental variables rather than fixing them.

Domain resolution. SerpAPI returns fully-qualified hostnames, whereas Tranco ranks registrable domains. Matching hostnames directly scores en.wikipedia.org as distinct from wikipedia.org, and since Wikipedia is the most frequently returned domain in our sample, the choice is not marginal. We report both the naive host-level rule and resolution to eTLD+1 against the Public Suffix List, using the snapshot bundled with a pinned tldextract so that normalization does not drift as the list is updated upstream. The choice leaves the frame itself untouched at 1,000 rows,

Tranco already being a list of registrable domains; it changes only which SERP results are attributed to a frame domain.

Rank cutoff. Presence anywhere in the returned results is a weaker claim than presence in the top three. We evaluate both.

The query subset controls only the labels; features are computed once and held fixed, so performance changes across budgets are attributable to supervision alone.

C. Evaluation Protocol

All `serp_*` columns are excluded from every design matrix, since they are functions of the label.

At the prevalences studied here, one to eleven percent, threshold metrics are close to uninformative: a classifier predicting the majority class everywhere achieves high accuracy and an F1 of zero, which describes the prevalence rather than the model. We report PR-AUC as the primary metric, expressed as *lift* (PR-AUC over prevalence, since a random ranker scores exactly prevalence). ROC-AUC appears alongside because prior work reports it [16], [17], and Section IV shows the two disagreeing. F1 is

quoted at the optimal threshold, with the fixed-0.5 value retained only to size the artefact.

Every number is a mean over five independent five-fold stratified partitions, reported with the standard deviation across partitions. Partitions are seeded identically across configurations, so ablation comparisons are paired [21], [22].

The popularity control. Popular domains appear in search results more often, which makes rank a mechanically strong predictor of the label. Any claim that fused features carry signal must therefore clear a baseline that ranks domains by Tranco position alone and fits nothing. We report this baseline throughout, and a label-permutation null establishing the PRAUC the fused model reaches on shuffled labels.

Four classifier families are compared: ℓ_2 -regularised logistic regression, random forest, XGBoost and LightGBM [14], all with balanced class weighting, implemented with scikitlearn [15]. LightGBM is used for the ablations.

IV. RESULTS

We collected 500 Google result pages in August 2026, yielding 4,061 organic results over 2,185 distinct hostnames, with no failed requests. Under eTLD+1 resolution these induce 111 positives in the 1,000-domain frame, a prevalence of 11.1%; under host-level matching, 77 (7.7%).

A. Budget Governs Yield, Learnability and Stability (O1)

Figure 2 reports the budget sweep. Supervision yield follows Heaps' law closely, $V = 6.21K^{0.470}$ with $R^2 = 0.993$ and residual RMSE of 2.7 domains. The exponent falls inside the 0.4–0.6 band reported for vocabulary growth in text corpora [1], which is not obviously expected: one process samples word types from documents, the other samples domain occurrences from result pages.

Distinct domains grow far faster, at $K^{0.865}$ ($R^2 > 0.999$), and show no sign of levelling off anywhere in our range. Put the two exponents side by side and the mechanism is visible. Sampling the open web is sampling something effectively unbounded, so discovery stays close to linear; the frame, by contrast, is 1,000 fixed domains being steadily used up. The square-root return belongs to the frame, not to the web.

Learnability follows. Mean PR-AUC climbs from 0.051 at $K = 10$ to 0.329 at $K = 500$, while the popularity baseline manages only 0.089 to 0.180 over the same span. The two therefore pull apart. At $K = 10$ the fused model is *behind* by 0.039; the curves cross near $K \approx 25$; by $K = 500$ the gap has opened to 0.149 in the other direction.

Stability of the interpretation lags further still[18]. Kendall τ between the top-20 feature-importance ranking at budget K and at full budget reaches only 0.45 at $K = 10$ and 0.62 at $K = 100$, the budget a free API tier affords. Conclusions about *which* signals matter are therefore substantially less reliable than conclusions about whether any signal exists, at every budget we examined.

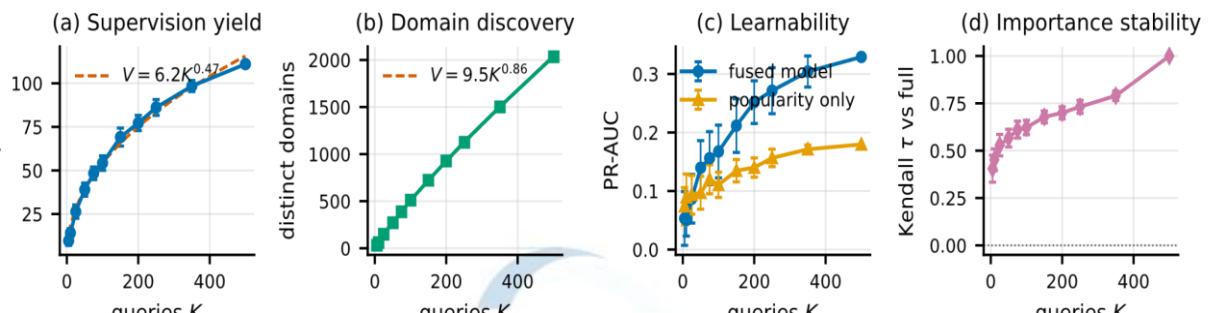


Fig. 2. Effect of query budget, holding features fixed and rebuilding only the labels. Points are means over 30 independent query draws per budget; bars are one standard deviation. (a) Positives grow as $K^{0.47}$. (b) Distinct domains observed grow as $K^{0.87}$. (c) The fused model trails the popularity ranker below $K \approx 25$ and leads thereafter. (d) Agreement between the feature-importance ranking at budget K and at full budget.

TABLE I

SERP-PRESENCE CLASSIFICATION AT $K=500$, FIVE REPEATED FIVE-FOLD PARTITIONS. F_1^* IS TAKEN AT THE OPTIMAL THRESHOLD AND $F_1^{0.5}$ AT A FIXED 0.5. *Prior* IS THE SAME MODEL ON THE 100-QUERY SAMPLE (35 POSITIVES) REPORTED PREVIOUSLY, FOR WHICH NO DISPERSION WAS AVAILABLE.

Model	PR-AUC	SD	ROC-AUC	F_1^*	$F_1^{0.5}$	Prior
Logistic regression	0.238	0.008	0.762	0.367	0.351	0.104
Random forest	0.310	0.006	0.799	0.371	0.042	0.145
XGBoost	0.321	0.016	0.804	0.389	0.177	0.181
LightGBM	0.329	0.012	0.813	0.403	0.254	0.203
Popularity baseline	0.180	0.000	0.567	0.216	0.000	—

B. Classifier Families

Table I sets four classifier families against each other at full budget. Their ordering is unchanged from an earlier 100query run of this pipeline, logistic regression < random forest < XGBoost < LightGBM, so the extra supervision did not rearrange anything. It lifted everything, by +0.126 to +0.165 PR-AUC, which is enough to change the verdict on two of them: both previously scored under the popularity ranker and now sit above it.

The F_1 columns show why threshold choice is not a detail. At a fixed 0.5 the random forest

scores 0.042 against its own optimum of 0.371, while logistic regression barely moves. The artefact is not a constant offset: it reorders the models, and it is what produced a reported F_1 of zero for the second-best classifier in that earlier run.

C. Source Ablation (O2)

The fused representation[19] reaches PR-AUC 0.329 ± 0.012 at full budget, or $2.96\times$ prevalence, against 0.180 for the popularity ranker. Shuffling the labels 100 times puts the

null at mean 0.118 with a 95th percentile of 0.141, so $p = 0.0099$.

That advantage is unevenly distributed, as Figure 3 makes plain. Take Tranco rank out and PR-AUC goes *up*, to 0.336; take the live crawl out and it barely moves, to 0.330. Both land within

noise of the full representation or above it. CrUX and Common Crawl are a different story, costing 0.051 and 0.056 when removed. Run each source alone and only Common Crawl (0.244) manages to beat the popularity baseline. CrUX

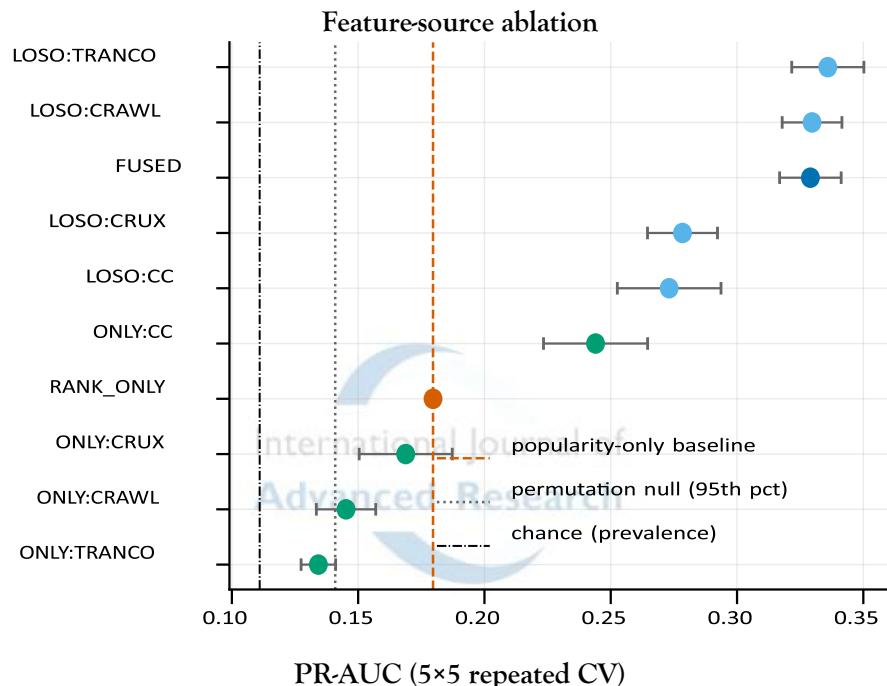


Fig. 3. Feature-source ablation at $K = 500$. LOSO: X removes source X ; ONLY: X retains only X . RANK_ONLY scores domains by Tranco position with no model fitted.

reaches 0.169, the live crawl 0.145, and rank fitted as a lone feature 0.134.

So the effect is a specific pairing, Common Crawl with CrUX, reaching 0.336 where neither source alone clears 0.25. One oddity is worth recording because it is easy to reproduce by accident: rank fitted as a feature (0.134) does *worse* than the same rank used as an unfitted score (0.180). Cross-validating a monotone signal can destroy information rather than extract it.

D. Protocol Ablations

Table II holds features and folds fixed and varies one protocol choice at a time. Switching domain resolution changes the positive count by 44%, from 111 to 77, and the measured advantage over popularity from +0.149 to +0.216. Tightening the cutoff to top-3 pulls both down. The four label-construction cells span +0.114 to +0.216 in all. That is close to a factor of two, produced by decisions a methods section normally disposes of in one clause.

Threshold policy behaves as expected under imbalance. LightGBM scores $F_1 = 0.254$ at a fixed 0.5 threshold and

TABLE II PROTOCOL ABLATIONS. EVERY ROW USES IDENTICAL FEATURES AND FOLDS; ONLY THE STATED PROTOCOL CHOICE VARIES.

Ablation	Setting	Positives	PR-AUC	SD	Note
Label construction	etld1 / top10	111	0.329	0.012	rank-only 0.180
Label construction	etld1 / top3	75	0.292	0.021	rank-only 0.169
Label construction	host / top10	77	0.367	0.049	rank-only 0.151

Label construction	host / top3	53	0.249	0.026	rank-only 0.135
Missing values	native	-	0.329	0.012	
Missing values	median	-	0.302	0.019	
Missing values	indicator	-	0.306	0.017	
Threshold	lightgbm	-	0.329	-	F1@0.5=0.254, F1@opt=0.403
Threshold	logreg	-	0.238	-	F1@0.5=0.351, F1@opt=0.367

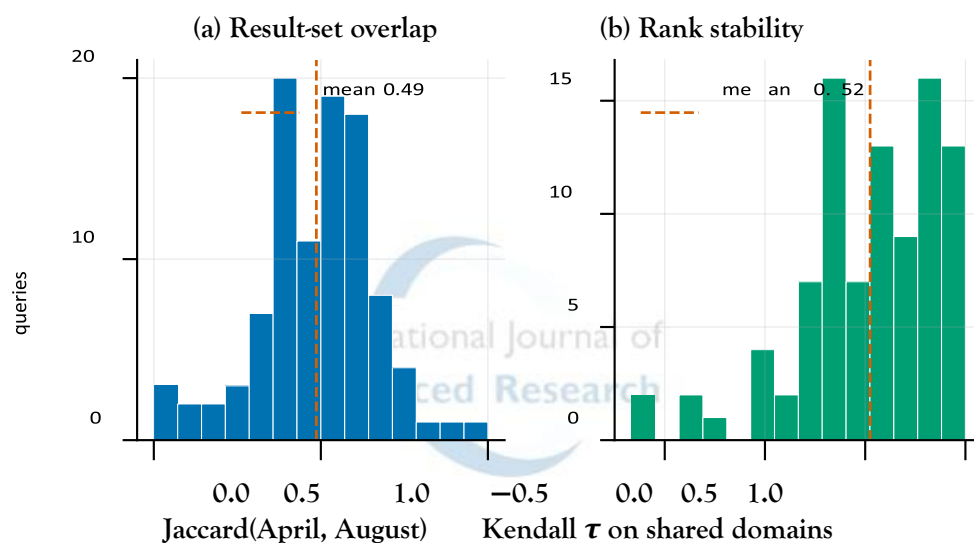


Fig. 4. Four-month drift over the 100 queries present in both collections. (a) Jaccard overlap of the returned domain sets. (b) Kendall τ over domains ranked in both snapshots.

0.403 at the optimal threshold of 0.021, which is nowhere near 0.5. The same comparison on the April sample, at 4.8% prevalence, gave a ratio of 3.5 $\times$ rather than 1.6 $\times$: the fixedthreshold figure degrades precisely as budgets shrink.

Missing-value policy is the one choice here that does not much matter. Native NaN routing reaches 0.329 ± 0.012 , median imputation 0.302 ± 0.019 , and a missingness indicator 0.306 ± 0.017 . Native is modestly ahead and no less stable, which suggests absence in CrUX and Common Crawl coverage says little the observed values do not.

E. Supervision Decays (O3)

All 100 April queries were re-issued in the August collection under an identical protocol, isolating four months of drift. Figure 4 reports the distributions. Mean Jaccard overlap between the two result sets is 0.487 (median 0.500), rising

only to 0.546 when restricted to the top three positions. Kendall τ over domains present in both snapshots is 0.525. Each query gained 2.6 and lost 3.5 domains on average.

At the frame level the effect is milder but still material. Of 48 positives in April and 45 in August, 38 persist, 10 are lost and 7 are new, a label-set Jaccard of 0.691. Roughly a third of the supervision purchased in April no longer described the August result pages.

F. A Budget Recommendation (O3)

Inverting the fitted yield curve, $K = (P/6.21)^{1/0.470}$, converts a target positive count into a query budget. Reaching 15% prevalence on a 1,000-domain frame requires roughly 870 queries; 20% requires about 1,600; 25% about 2,600. Against the free tier that most exploratory work runs on, these are one to two orders of magnitude larger.

Two thresholds are more useful than the curve itself. Below roughly 25 queries, a fused model is outperformed by sorting on popularity, so any positive claim about feature value is

unsupported. Below roughly 100, feature-importance rankings agree with the full-budget ordering at $\tau < 0.65$, so claims about which signals matter should not be made. Both bounds sit above what free API tiers provide.

V. DISCUSSION

A. Labels, Not Rows, Are the Binding Constraint

Neither the 1,000 domains nor the 102 features are the constraint here. The constraint is how few of those domains are ever seen in a result page at all, and that count climbs only $\sqrt{K}$ as K . Adding *domains* buys nothing, because every row already carries its features. Adding *queries* buys something, slowly. Faced with a weak model most of us reach for more features or more rows first, which leaves the actual bottleneck untouched.

B. Why the Popularity Baseline Matters

At eleven percent prevalence, almost anything beats chance. Sorting domains by Tranco rank is the comparison that carries information, since it costs nothing, fits no parameters, and is already entangled with the label through traffic. Three of our four single-source models clear chance and still land *below* it. Each would have been written up as a positive result had we quoted lift over prevalence and stopped there.

The two metrics also disagree about which configuration wins, as Saito and Rehmsmeier describe [6]. ONLY:CRUX reaches ROC-AUC 0.833 under host resolution while its PRAUC sits at 0.285: it orders the negative majority well without surfacing positives near the top, which is the only behaviour a shortlisting task consumes.

C. Budget Determines the Conclusion, Not Just the Confidence

A smaller sample is normally understood to widen error bars around a stable estimate. That is not what happens here. The sign of the fused-versus-popularity comparison changes: negative below $K \approx 25$, positive above it, and still growing at $K = 500$. Two studies following identical protocols on different API tiers would publish contradictory conclusions, and both would be reporting their data faithfully.

This bears on how such results should be read. A singlebudget experiment does not license a claim about which features matter; it licenses a claim about which features matter *at that budget*. Our own earlier reading of these data at $K = 100$ found no advantage for fusion, and it took a fivefold larger budget to show that conclusion was an artefact of label scarcity.

D. Protocol Choices Are Results

How SERP hostnames map onto frame domains is normally compressed into a phrase, if stated at all. It moves the measured advantage of fusion by almost a factor of two, from +0.114 to +0.216 depending on resolution and rank cutoff, and it changes the positive count by 44% (111 against 77).

The decision threshold behaves similarly, and interacts with budget in a way worth naming: the gap between F_1 at 0.5 and F_1 at the optimal threshold is 1.6 $\times$ at 11% prevalence but 3.5 $\times$ at 4.8%. The metric is least trustworthy exactly where budgets are smallest, which is where it is most often reported.

E. Which Sources Are Worth Collecting

The ablation carries a practical implication that the aggregate comparison hides. Common Crawl and CrUX are complementary: 0.244 and 0.169 in isolation, 0.336 together. Tranco rank and the live crawl are not. Dropping rank from the fused representation raises PR-AUC from 0.329 to 0.336, and dropping the crawl leaves it unchanged within one standard deviation.

The live crawl was the most expensive source here, requiring per-domain rate limiting, robots compliance and retry handling, and it still reached only 53.6% of the frame. Two static bulk sources needing no crawling deliver the entire measured effect.

F. Limitations

Query budget. Our largest budget remains modest relative to the distribution being sampled. The exponent is estimated over budgets spanning a hundredfold range and fits well inside it, but projecting an order of magnitude beyond is extrapolation. One out-of-range check is available: a fit to the April sample alone, which never saw budgets above $K = 100$,

projected 107 positives at $K = 500$ against 111 observed. That is reassuring at $5\times$ and says nothing about $50\times$.

Frame composition. The Tranco top-1,000 over-represents infrastructure and platform domains relative to the general web [3], and several of the most frequently ranked domains in our SERP sample lie outside it entirely. Conclusions about which features predict presence are conditional on that frame.

Locale and query distribution. All queries went to Google in the US English locale, and MS MARCO queries derive from Bing logs and skew informational. A different locale or a commercially skewed query set would induce a different domain distribution and plausibly a different exponent.

VI. CONCLUSION

Treating search result pages as ground truth makes the query budget a sampling instrument, and we measured what that instrument does. Labelled positives return the square root of spend, $6.21K^{0.47}$, because a popularity frame saturates while the web behind it does not. Reaching 15% label prevalence on a thousand-domain frame needs on the order of 870 queries, well beyond the free tiers on which exploratory work is usually done.

That costs more than precision. Below about 25 queries a fused representation loses to sorting domains by popularity; above it the same representation wins, by a margin that keeps widening out to 500. We landed on both answers ourselves, four months apart, running the same code at different budgets.

A methods section records the model, the features and the folds. It does not record where on this curve the study was standing, which is what decided the answer.

Fusion helps narrowly. Common Crawl and Chrome UX Report carry all of the measured effect[20]. Tranco rank contributes nothing, and neither does a live five-page crawl, which is awkward given that the crawl dominated our collection cost. Two labelling choices normally given a clause, resolving hostnames to registrable domains and picking a rank cutoff, move the measured advantage by nearly a factor of two. Supervision also spoils: a third of the positive

labels bought in April had stopped describing the August result pages.

Three things would help. Report the query budget as an experimental parameter rather than a footnote. Benchmark against popularity, not against chance. State the domain resolution rule. Without them a reader cannot tell a genuine disagreement in this literature from two groups on different API tiers.

REFERENCES

- L. Egghe, "Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 702–709, 2007.
- V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, 2019.
- Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann et al., "A long way to the top: Significance, structure, and stability of internet top lists," in *Proceedings of the Internet Measurement Conference (IMC)*, 2018, pp. 478–493.
- A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Re, "Snorkel: Rapid training data creation with weak supervision," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.
- J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 233–240.
- T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.

- A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 265–270, 1984.
- T. Viering and M. Loog, "The shape of learning curves: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7799–7819, 2023.
- Google, "Chrome UX report," <https://developer.chrome.com/docs/crux>, 2026, accessed 4 August 2026.
- Google Chrome Developers, "Advancing Interaction to Next Paint," <https://web.dev/blog/inp-cwv>, 2024, INP replaced FID as a Core Web Vital in March 2024; accessed 4 August 2026.
- Common Crawl Foundation, "Common crawl: Open repository of web crawl data," <https://commoncrawl.org>, 2026, CDX index queried August 2026; accessed 4 August 2026.
- SerpApi, "SerpApi: Google search API," <https://serpapi.com>, 2026, accessed 4 August 2026.
- T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary et al., "MS MARCO: A human generated machine reading comprehension dataset," in *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo@NIPS)*, ser. CEUR Workshop Proceedings, vol. 1773, Barcelona, Spain, 2016.
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 3146–3154.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- Ahmad, J., F. Tauseef, and Z. Akbar. "Predictive analytics for AI-assisted patient no-show management and clinic revenue optimization: a simulation-based research." *Migration Letters* 21, no. S13 (2024): 1901-1924.
- Tauseef, Fatima, Ahmad Jamal, Fahad Naseer, Mahzabin Mahmud, Md Mishal Mahmood, Md Ragybul Islam, and Samia Ara Chowdhury. "MACHINE LEARNING MODELS FOR PREDICTING PATIENT OUTCOMES IN INTENSIVE CARE UNITS: A CASE STUDY IN US HOSPITALS." *Cuestiones de Fisioterapia* 55, no. 1 (2026): 88-105.
- Ahona, M. H., Jamal, A., Tauseef, F., Mahmud, M., & Indika, M. H. (2026). AI-driven discovery of novel biomarkers for early cardiovascular disease detection. *Well Testing Journal*, 35(S2), 1–46. <https://doi.org/10.66669/WT.v35iS2.287>
- Jamal, A., Akbar, Z., Akbar, S., Niaz, S., & Tauseef, F. (2022). Predicting human-GenAI collaboration effectiveness: A machine learning investigation of skill configurations, trust, and work design. *Migration Letters*, 19(S8), 2303–2324. <https://doi.org/10.53555/ml.v19iS8.12298>
- Jamal, Ahmad, Fatima Tauseef, Fahad Naseer, and Jawad Ahmad. "BUSINESS ANALYTICS FOR HEALTHCARE COST OPTIMIZATION: A DATA-DRIVEN STUDY ON IMPROVING FINANCIAL SUSTAINABILITY IN HEALTHCARE SYSTEMS." *Spectrum of Engineering Sciences* (2024): 723-736. [doi:zenodo.org/records/20488115](https://doi.org/10.24018/engsci.2024.723-736).
- Nasim, F., Yousaf, M.A., Masood, S., Jaffar, A. and Rashid, M., 2023. Data-driven probabilistic S for batsman performance prediction in a cricket match. *INTELLIGENT AUTOMATION AND SOFT COMPUTING*, 36(3), pp.2865-2877.
- Mehdi, M., Nasim, F. and Munir, M.Q., 2024. Comparative Risk Analysis and Price Prediction of Corporate Shares Using Deep Learning Models like LSTM and Machine Learning Models. *Journal of Computing & Biomedical Informatics*, 7(02).