Half of B2B SaaS Brands Are Absent from Open-Weight Parameters, and One in Five of the Rest Is Misfiled

A parametric audit of 1,000 public B2B SaaS brands in Gemma 4 12B, with a 79-variable analysis of what puts a brand there

Ben Moore
Discovered Labs
https://discoveredlabs.com
ben@discoveredlabs.com

Liam Dunne
Discovered Labs
https://discoveredlabs.com
liam@discoveredlabs.com

August 30, 2026

Abstract

Brand visibility in language models is usually measured behaviourally, by sampling what an assistant says. That conflates what a model has stored in its weights with what it retrieved at answer time. We take the parametric reading directly on an open-weight model, auditing 1,000 public B2B SaaS brands across 333 categories in Gemma 4 12B. Unprompted category recall names 51.0% of the cohort. Of the brands the model will commit to placing, 194 of 981 (19.8%) are filed under the wrong product category, and only 19 produce no preference at all, so the model almost always commits whether or not it knows. Accuracy is steeply concentrated in the decision margin: the 547 brands above a 0.6 margin err 3.5% of the time, while the 318 in the two lowest bands err at 39.9% and 52.0%. Recall rose 11.3 points across three Gemma generations and then stopped, with Gemma 4 at 51.0% against Gemma 3 at 51.1% on identical protocol. Regressing 79 brand-level variables on both outcomes, holding a fixed confounder block of crawl-graph rank and referring domains, separates the two: a technical-authority composite predicts being named (5.37 points per standard deviation, $p_{adj} = 0.006$) but not being described correctly (-2.77), while an owned-properties composite predicts being described correctly (5.92, $p_{adj} < 0.001$) and not being named (1.36, n.s.). Model sentiment is unrelated to customer satisfaction ($r = 0.03$, $n = 921$). We argue that recall and correctness are separate failure modes with separate levers, and that the parametric layer must be measured on open weights rather than inferred from behaviour.

Keywords: parametric knowledge; language model probing; brand visibility; membership inference; observational study.

Contributions

1. **A parametric cohort audit.** 1,000 public B2B SaaS brands across 333 categories, read from the weights of an open-weight model rather than from generated text, so retrieval cannot enter the measurement.
2. **Correctness as a margin function.** Cohort accuracy decomposes into a steep curve: 3.5% error above a 0.6 decision margin against 52.0% in the lowest band, with the margin invisible to a reader.
3. **A generational trajectory.** Identical protocol across four Gemma generations gives an 11.3-point climb followed by a plateau.
4. **Disjoint predictors.** 79 variables under a fixed confounder block separate being named from being described correctly, with technical surfaces buying the first and owned reference material the second.
5. **A null with commercial force.** Model sentiment is uncorrelated with real customer ratings ($r = 0.03$, $n = 921$), and the weights hold essentially nothing about pricing.

1 Introduction

Buyers increasingly begin product research inside a language model rather than a search engine, which makes a commercial question out of an interpretability one: what does a model hold about a company, and where did it come from?

The available answer is almost always behavioural. A prompt is issued to a hosted assistant and the response is scored for whether the brand appears and whether the description is right. That procedure measures a pipeline, not a model. A hosted assis-

tant retrieves documents, reranks them, and conditions on them before generating, so a brand can be named because a crawler fetched its homepage seconds earlier rather than because anything about it survived training. The two are commercially different. Retrieved presence is rented and moves within days; parametric presence is owned, changes only when a model is retrained, and then persists across every downstream product built on those weights.

Separating them requires a model whose weights can be read. We therefore run the audit on Gemma 4 12B, an open-weight model, and treat closed models as a behavioural cross-check rather than as the measurement.

Contributions. In this paper, we (i) audit 1,000 public B2B SaaS brands across 333 categories for unprompted category recall and for stored category correctness, taking both readings from the weights rather than from generated text; (ii) show that correctness is not a single cohort number but a steep function of the decision margin, and that the margin is invisible to the reader of an answer; (iii) track the same protocol across four Gemma generations and find an eleven-point climb followed by a plateau; (iv) regress 79 brand-level variables on both outcomes under a fixed confounder block and show that recall and correctness have largely disjoint predictors; and (v) report that model sentiment is uncorrelated with real customer satisfaction, so it measures how a company is written about rather than how it performs.

2 Related work

Reading facts out of weights. A body of work probes language models for stored relational knowledge by scoring candidate completions rather than sampling them, beginning with cloze-style factual probes (Petroni et al., 2019) and continuing through work that localises and edits specific factual associations (Meng et al., 2022). Our category probe is in this tradition: the model ranks a fixed set of candidate sentences and we read which one the weights prefer, so no text is generated and nothing can be hallucinated into the measurement.

Membership inference and training-data attribution. Whether a particular document was in a model's training set has been attacked through loss-based and calibrated membership inference, most relevantly the Min-K% family and its normalised variants. We use a reference-normalised Min-K%++ detector for the training-data survey in Section 5, validated against documents with provable membership.

Brand and entity visibility. Commercial measurement of brand presence in AI answers is almost entirely behavioural, scoring hosted assistant responses. We are not aware of prior work that takes the parametric reading at cohort scale and contrasts it against the behavioural one on the same brands, which is the gap this paper addresses. Where prior commercial studies report retrieval-time citation drivers, our results are complementary rather than competing: they describe a different layer of the same pipeline.

What is not established. We found no prior published cohort-scale estimate of misfiling rate, of the margin–accuracy relationship, or of the generational trajectory of brand recall in an open-weight family, so those three results have no baseline to be compared against and should be read as first estimates rather than as replications.

3 Data

Cohort. The unit of analysis is a brand. We assembled 1,000 public B2B SaaS companies spanning 333 product categories, from category incumbents down to seed-stage companies. Selection is public-company only; no client or private data enters the cohort, and no brand was included or excluded on the basis of any measurement reported here. Each brand carries a verified product category, drawn from public sources and used as ground truth for the fact probe.

Models. All parametric readings are taken on Gemma 4 12B. The generational comparison in Section 5 adds Gemma 3 12B, Gemma 2 9B and Gemma 7B, each run on the identical cohort, prompt set and matcher. The behavioural cross-check uses GPT-5.6 Luna, Claude Haiku 4.5 and Gemini 3.5 Flash Lite through their public APIs.

Brand-level variables. For the analysis in Section 5 we constructed 79 brand-level variables, grouped into four families used for multiple-comparison correction:

- **Coverage volume** — mention counts and audience size across 17 measured surfaces, plus Wikipedia citations of the brand's domain, snippet length, title share and corpus age statistics.
- **Coverage language** — for each surface, the share of snippets mentioning the brand that also state its product category. This is the *category co-occurrence* family and is a property of what coverage says rather than of how much of it exists.

- **Owned properties** — presence and depth of a help centre, documentation passages, pricing disclosure, free-tier and self-serve status, an integrations directory, and the model's measured familiarity with the brand's own site.
- **Namespace** — whether the brand name is an ordinary dictionary word, and how crowded its product category is.

Confounder block. Two authority variables, rank in the Common Crawl host graph and count of referring domains, are carried as a fixed confounder block in every adjusted model rather than being reported as findings alongside the rest. They are strongly associated with both outcomes and with almost every other variable, so leaving them free would let company size masquerade as a content effect.

Availability. Per-variable sample sizes differ because not every brand has every surface. We report n with every estimate. The largest models run on $n = 998$; surface-specific models run on 855–909.

4 Methods

4.1 Unprompted category recall

For each of the 333 categories we issue two unbranded prompts of the form a buyer would use, for example `"the most widely used CRM software tools for businesses are"`, and sample six completions per prompt at $T = 0.7$. A brand counts as recalled if its name appears, word-boundary matched with suffix stripping and domain forms accepted, anywhere in its category's twelve completions. Word-boundary matching matters: an earlier substring matcher produced a false positive rate high enough to invert one brand's result entirely, by matching a short brand name inside longer ordinary words.

4.2 Stored category correctness

The fact probe is a ranking, not a generation. For each brand we construct five candidate sentences that differ only in the product category named, one true and four distractors, and score each by mean token log-probability. The brand's outcome is read off the ranking:

- **Consistent** — the true category ranks first.
- **Deviated** — a distractor ranks first.
- **Unaddressed** — no candidate separates by a meaningful margin.

The *margin* is the log-probability gap between the winner and the runner-up, and is the quantity that carries the accuracy structure in Section 5.

Two corrections were necessary and both changed the headline. First, distractors are resampled from the cohort's 333 categories constrained to the same word count as the true label, with character length as a tiebreak, so the shape of a label cannot carry the ranking. Second, sequences are right-padded, so the beginning-of-sequence token is never scored as a target. Under left padding the score depends on the longest option in the batch, which is a property of the batch rather than of the model; that scoring puts 663 brands in the consistent bucket against 787 under the corrected scoring, and we report only the corrected run.

4.3 Commercial facts

Three commercial facts, free-tier existence, published pricing and self-serve signup, plus integration partnerships, are tested with a calibrated design that holds the wording fixed and varies the brand. The statistic is AUC over the positive-minus-negative log-probability lean, which asks whether brands for which the claim is true lean higher than brands for which it is false. This cancels any constant preference for one phrasing over its negation, which an earlier per-brand ranking design had measured instead of brand knowledge. Both sides of every integration pair are cohort brands, so a probe cannot be won by recognising one name as real and the other as invented.

4.4 Sentiment

Sentiment is read as a linear direction from antonym contrast pairs, mean-pooled over the brand-name token span located by character offsets at layer -8, on twelve dimensions. Each dimension uses its own prompt frame, so raw scores are not comparable across dimensions; every reported quantity is therefore either a within-dimension comparison between brands or a z-score against the cohort on that dimension.

4.5 Association analysis

Each of the 79 variables is fitted in a separate logistic regression against each outcome, carrying the fixed confounder block, and reported as an average marginal effect in outcome points per standard deviation of the variable. The accuracy models carry one further control for the true label's rank position in the probe. Multiplicity is handled with the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) within variable family and outcome.

The six channel composites are built by z-scoring components and averaging, requiring at least two present components per brand, and are corrected across the twelve composite tests.

This is an observational design on a cohort we did not randomise, so every estimate below is an association under a stated adjustment set, not a causal effect.

5 Results

5.1 Half the cohort is never named

Unprompted category recall on Gemma 4 12B names 510 of 1,000 brands. The 490 that go unnamed are not shell companies: the list includes established products in categories the model discusses fluently. Recall varies sharply by category, from 8 of 12 brands surfaced in project management down to 2 of 11 in education technology, so a crowded category is not the same as a visible one and the incumbents' grip on the answer space is the relevant baseline for any individual brand.

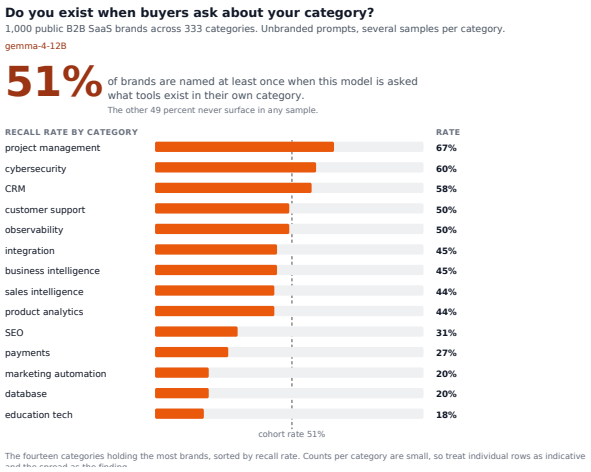

Figure 1: Unprompted recall on Gemma 4 12B, overall and by category. Half the cohort is absent from the model's answer space for its own category.

Reseeding the same prompts flips 13% of individual brands between named and not named, so a single reading of one brand is a noisy lower bound. Cohort-level contrasts are an order of magnitude above that floor and are the level at which we make claims.

5.2 The climb across generations has stopped

Table 1 runs the identical protocol across four Gemma generations. Recall rises 39.8 → 44.3 → 51.1 and then stops, with Gemma 4 landing at 51.0 against Gemma 3's 51.1 at the same parameter scale. The 11.3-point climb is far above the reseed noise floor; the −0.1 step is far below it. Read together, the imprint compounds across training cycles but is not handed out automatically, and this cohort's coverage bought three generations of gains before it stopped buying.

Table 1: Unprompted recall across four Gemma generations on the identical cohort, prompt set and matcher. The eleven-point climb is an order of magnitude larger than the 13% per-brand reseed flip rate; the plateau between the last two generations is not.

Generation	Model	Recall
Gemma 1	gemma-7b	39.8%
Gemma 2	gemma-2-9b	44.3%
Gemma 3	gemma-3-12b-pt	51.1%
Gemma 4	gemma-4-12B	51.0%

5.3 One brand in five is filed under the wrong category

Table 2 gives the stored-category outcomes. The model places 787 brands correctly and 194 incorrectly, and produces no preference at all for only 19. That last number is the uncomfortable one: a model that hedged when it did not know would be safer for a buyer, but this one commits on 98.1% of the cohort regardless.

Table 2: Stored category outcomes for 1,000 brands on Gemma 4 12B, under the corrected probe. The model commits on 98.1% of brands whether or not it knows.

Outcome	Brands	Share
Consistent (true category first)	787	78.7%
Deviated (a distractor first)	194	19.4%
Unaddressed (no separation)	19	1.9%
Median margin, consistent		0.83
Median margin, deviated		0.21

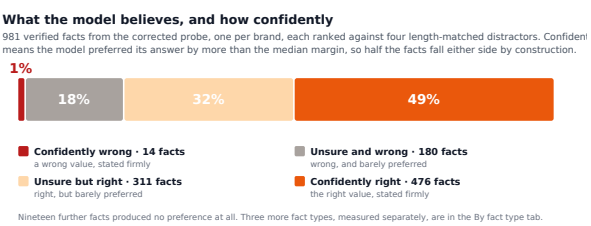

Figure 2: Outcomes crossed with confidence. Only 14 facts in 1,000 are confidently wrong, but a further 180 are wrong without signalling any doubt.

Deviations are held less firmly than correct answers, with a median margin of 0.21 against 0.83,

but the deviation distribution keeps a confident tail: 103 of the 194 still clear a 0.2 margin. Correctness is not safe either, since 84 of the 787 correct placements win by less than 0.2, with Sumo Logic holding observability by 0.020 and Conductor holding SEO by 0.024.

Being well known does not protect a brand. The model files Okta under CI/CD rather than identity, Square under card issuing rather than SMB commerce, Akamai under CI/CD rather than CDN, and HashiCorp under workflow orchestration rather than infrastructure security.

5.4 Accuracy is a function of the margin, not of the cohort

Table 3 shows that the cohort-wide error rate is an average over a steep curve. The 547 brands above a 0.6 margin err 3.5% of the time taken together and the highest band errs on none of its 66 brands, while the 318 brands in the two lowest bands err at 39.9% and 52.0%. Filtering on the margin therefore works: demanding 90% accuracy retains 742 brands and demanding 95% retains 609.

Table 3: Error rate by decision margin. Accuracy is not a cohort constant but a steep function of the margin, and the margin is not visible to a reader of the model's answer.

Margin band	Brands	Error rate
0.0–0.2	175	52.0%
0.2–0.4	143	39.9%
0.4–0.6	116	23.3%
0.6–0.9	175	6.3%
0.9–1.2	176	3.4%
1.2–1.6	130	1.5%
≥1.6	66	0.0%
All above 0.6	547	3.5%

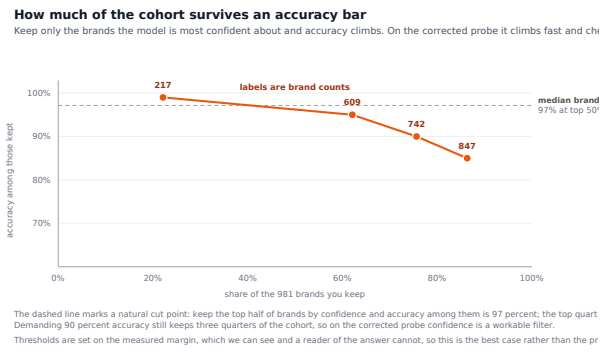

Figure 3: Coverage retained as the accuracy bar rises. Roughly a quarter of the cohort carries almost all of the error.

The practical caveat is that this threshold is set on a margin we can measure and a reader cannot. Nothing in a generated answer signals which band it came from, so the confident tail of the deviations reaches a buyer indistinguishable from the confident tail of the correct answers.

5.5 Deviations concentrate on magnet categories and on verticals

The 194 deviations are not uniformly distributed. They land on only 94 distinct categories, with ten absorbing 33% of the total, and the attractors are compact technical labels: CI/CD attracts 9 misfiled brands while holding 2 of its own, data lakehouse attracts 9 while holding 2, and web hosting attracts 6 while holding 1. These are labels the training text uses constantly and precisely, and they behave like sinks for brands the model cannot otherwise place. The pattern survives a synonym check: only 10 of the 194 share any content word with the true category.

Splitting by industry moves the deviation rate from 0% for brands selling IT and internal operations tools to 37% for those selling into regulated verticals such as health, education and government, with developer tools at 30% and finance and legal at 24%.

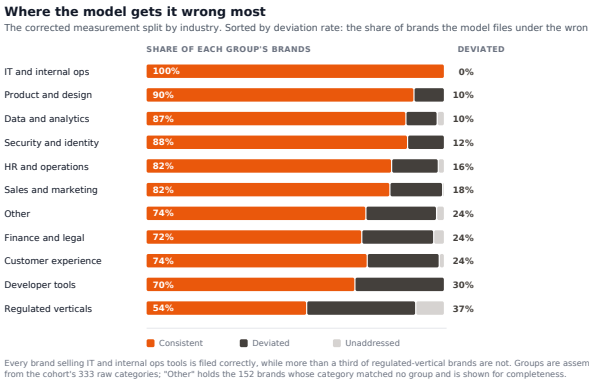

Figure 4: Deviation rate by industry group. The cohort-wide rate hides a seven-fold spread, and the worst groups are those where a confident error costs a buyer most.

5.6 The model knows what a company is, not how it sells

Table 4 separates fact types against their own baselines. Product category is right 78.7% of the time against a 20% chance baseline. Integration partnerships separate true from false pairs at AUC 0.614 on 2,400 probes, the best-held binary fact we tested. The three commercial facts are close to useless: free tier at 0.547, published pricing at 0.526 and self-serve at 0.450, which is below chance.

Table 4: Calibrated commercial-fact probes, AUC over the positive-minus-negative lean with wording held fixed and brand varied. Chance is 0.500. What a company is and who it connects to are held; how it sells is not.

Fact type	Probes	AUC
Product category (5-way)	1,000	78.7%†
Integration partner	2,400	0.614
Free tier exists	402	0.547
Published pricing	585	0.526
Self-serve signup	438	0.450

†Accuracy on a 5-way rank against a 20% chance baseline, not on the AUC scale.

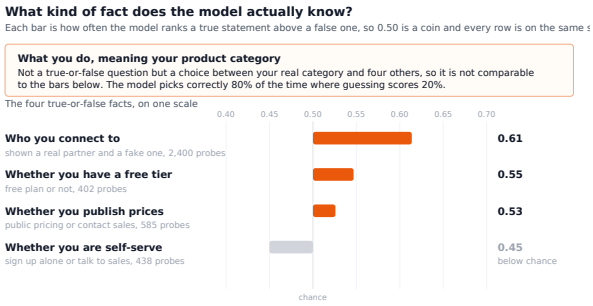

Figure 5: Each fact type against its own chance baseline.

The split tracks who else writes the fact down. An integration is documented by both parties, quoted in comparisons and listed in changelogs, so it accumulates independent restatements. Pricing appears once, on the vendor's own page, which is the surface the weights are least influenced by. Commercially this is the awkward result, because pricing is among the most common things a buyer asks an assistant and there is nothing in the weights to answer it from.

5.7 Recall and correctness have disjoint predictors

Table 5 is the central analytical result. Under a fixed confounder block of crawl-graph rank and referring domains, the technical composite predicts being named at 5.37 points per standard deviation ($p_{adj} = 0.006$) and is slightly *negative* on being described correctly (-2.77, $p_{adj} = 0.047$). The owned-properties composite is the mirror image: 5.92 points on being described correctly ($p_{adj} < 0.001$) and a null 1.36 on being named. Social prominence buys 9.75 points of recall and 0.21 points of accuracy, the clearest case of reach without comprehension in the study.

Two further variables are worth isolating. Category co-occurrence, the share of snippets naming a brand that also state its category, is the one content

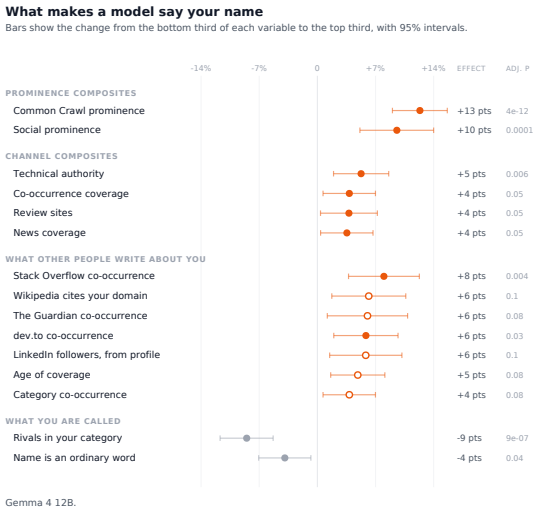

Figure 6: Variables that survive correction on being named.

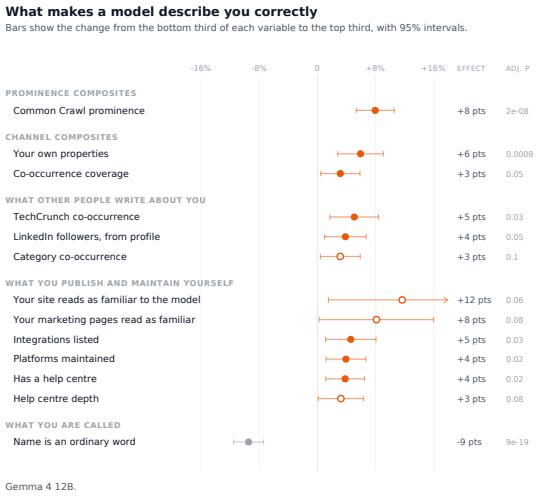

Figure 7: Variables that survive correction on being described correctly. The two lists barely overlap.

property with independent signal on recall, at 3.92 points per standard deviation, and it is bounded and at the edge of significance under the strictest correction. And an ordinary-word brand name is the single largest namespace effect, costing 14.1 points of recall and 32.0 points of accuracy unadjusted, which is a liability a marketer cannot fix after the fact.

The mechanism behind the technical/owned split is visible in the coverage language. A TechCrunch snippet states the brand's category 33.8% of the time and a Forbes snippet 28.2%, against 14.4% for GitHub and 9.8% for Stack Overflow. A dependency manifest records that a tool exists without ever saying what it is for, which is why developer surfaces move recall while doing comparatively little for correctness.

Table 5: Channel composites against both outcomes, in outcome points per standard deviation, adjusted for crawl-graph rank and referring domains and corrected across the twelve tests. The two outcomes have largely disjoint predictors: technical authority buys being named and slightly costs being described correctly, owned properties buy the reverse, and social prominence buys recall alone.

	Recall				Category accuracy			
Composite	n	pts/SD	95% CI	p_{adj}	n	pts/SD	95% CI	p_{adj}
Crawl prominence	885	12.58	[9.21, 15.94]	<0.001	869	7.93	[5.34, 10.53]	<0.001
Social prominence	909	9.75	[5.22, 14.27]	<0.001	892	0.21	[-2.43, 2.85]	0.911
Technical	885	5.37	[1.99, 8.76]	0.006	869	-2.77	[-5.22, -0.32]	0.047
Explanation	885	3.92	[0.70, 7.13]	0.045	869	3.15	[0.47, 5.84]	0.047
Review sites	884	3.87	[0.39, 7.36]	0.047	868	-0.80	[-3.20, 1.60]	0.586
News	885	3.62	[0.40, 6.83]	0.047	869	0.14	[-2.35, 2.63]	0.911
Owned properties	998	1.36	[-2.16, 4.87]	0.574	979	5.92	[2.79, 9.05]	<0.001
Social	871	1.29	[-2.19, 4.77]	0.574	855	-1.83	[-4.45, 0.79]	0.248

5.8 Sentiment is about coverage, not satisfaction

Star ratings and review counts from G2 and Trustpilot were available for 921 brands. The correlation between a brand's real customer rating and the sentiment held in the weights is $r = 0.030$ ($p = 0.36$), which is nothing. What the model absorbed is how a company is written about.

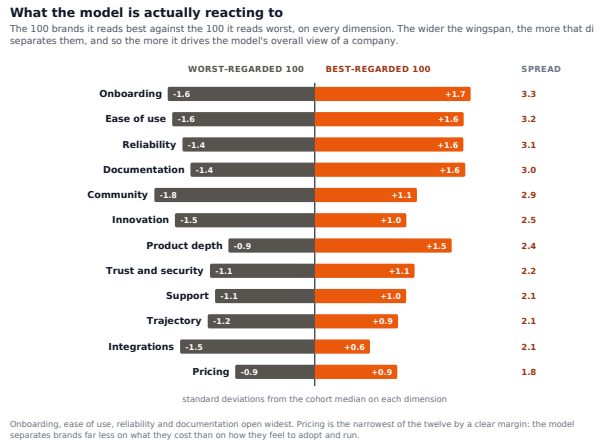

Figure 8: The 100 best-regarded brands against the 100 worst, on each of the twelve dimensions.

Ranking brands by their mean z-score across the twelve dimensions, onboarding separates the best hundred from the worst by 3.31 standard deviations, ease of use by 3.18, reliability by 3.07 and documentation by 3.00, while pricing value separates them by 1.77, the narrowest of the twelve. The dimensions are not independent: the median pairwise correlation is 0.517 and the first principal component accounts for 54.8% of the variance, with documentation and onboarding moving together at 0.889. Roughly 45% of the variance nonetheless sits outside that first factor, so the per-dimension readings carry real information even though they do not move independently of the headline.

5.9 What survives the training pipeline

The training-data survey, run with a reference-normalised Min-K%++ detector validated against documents with provable membership, finds a consistent shape preference in what reaches the weights. Scored on 1,477 passages, code and technical material leads at +1.82, structured lists at +1.12 and numeric material at +0.79, while personal opinion sits at −0.52, flowing prose at −0.61 and question-and-answer format at −0.71.

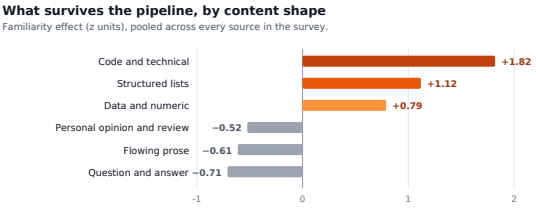

Figure 9: Detector score by content shape. Structure and specificity survive the admission pipeline; conversational and opinion formats do not.

Because these models are heavily distilled, the survey establishes what did make it in rather than what could not: absence in the detector is weak evidence of absence in training, not proof of it.

5.10 Behavioural cross-check

None of the closed models return the token probabilities a parametric reading requires, so they can only be asked. Running the same 1,000 brands through three of them, unprompted recall rises from 51.0% on Gemma 4 to 60.2% on Claude Haiku 4.5, 73.3% on Gemini 3.5 Flash Lite and 78.5% on GPT-5.6 Luna.

The larger difference is in whether a model declines. As a rate of being wrong when it commits, Gemma 4 sits at 19.8% against roughly 2% for Claude, 0.8% for Gemini and 0.4% for GPT, and the mechanism is abstention rather than knowledge: GPT declines to place 287 brands while Gemini declines on 8.

This comparison flatters the closed models and should be read that way. Gemma ranks five sentences with no options in front of it and no way to abstain, while the closed models see the candidate list, can reason over it, and are offered an explicit "I do not know". The recall figures are not like for like either, since Gemma was sampled twelve times per category against four for the closed models, which makes its lower recall the more striking of the two.

6 Discussion

Recall and correctness are separate problems with separate levers. The composite table is the practical core of the paper. A brand that is absent from the model's answer space and a brand that is present but misfiled look similar behaviourally, in that neither gets a correct mention, but they respond to different work. Technical-surface coverage moves the first and does nothing for the second; the brand's own reference material moves the second and does nothing for the first. Diagnosis therefore has to precede tactics, and a programme that measures only one of the two outcomes will report progress on work that is not addressing the actual failure.

Confidence is not observable at the point of use. The margin structure in Table 3 would be reassuring if it were legible. A reader who could see the margin could discount low-margin answers and be right 96.5% of the time on the rest. No such signal reaches them. The practical consequence is that "the model stated it confidently" carries no information about correctness for any individual brand, and 103 of the 194 wrong answers arrive above the same margin threshold that most correct ones clear.

Almost nothing is unaddressed, which is worse than it sounds. Only 19 brands produce no preference. A model that abstained under uncertainty would convert misfilings into visible gaps, which a brand could then work on and a buyer could discount. Instead the failure mode is a confident, specific, wrong category. The closed-model comparison shows this is a design choice rather than a fact about knowledge: GPT declines on 287 brands and is consequently wrong on 3.

Absence of a lever is itself commercially useful. That the weights hold essentially nothing about pricing, free tiers or self-serve signup is a negative result with a positive implication. Any assistant answer about a vendor's pricing is produced at retrieval time, so a crawlable, plainly worded pricing page controls that answer outright and no competitor can out-imprint it. The corresponding warning is that the reverse holds for category: a misfiling sits in the weights and will not be corrected by a page published today.

Two clocks. The generational result implies that parametric work pays on a slower cycle than retrieval work. Alignment layers update close to continuously, continued pretraining lands on a monthly-to-quarterly cycle, and genuinely new base models arrive every few months. Coverage published now reaches the weights at the third of those, which makes this a compounding programme rather than one with weekly readouts, and makes the Gemma 3 to 4 plateau an argument for deliberate work rather than for waiting.

Sentiment measures the corpus, not the product. An r of 0.03 against real customer ratings means the twelve-dimension readings should never be presented as a satisfaction proxy. They describe how a company is written about, which is a legitimate object of measurement for a marketer and a misleading one for anyone reading it as a verdict on the product.

7 Limitations

7.1 Critical

- **Observational, not causal.** No variable was manipulated. Every estimate in Table 5 is an association under a stated adjustment set, and the confounder block cannot rule out unmeasured company-level characteristics that drive both coverage and imprint.
- **One model family for the parametric readings.** All weight-level results are Gemma. The generational series shows the protocol transfers within the family, but nothing here establishes that the technical/owned split holds in a differently-trained model, and the closed models cannot be tested this way because they do not expose log-probabilities.
- **Recall is a lower bound.** Twelve samples at one temperature will miss genuinely known names. A miss is evidence of weak association, not proof of absence, and per-brand results carry a 13% reseed flip rate.

- **Distillation limits the training-data survey.** The detector establishes what did reach the weights. Because these models are heavily distilled, material can influence a model without appearing in the survey, so absence is not evidence of exclusion.

7.2 Notable

- **Category ground truth is a single label.** Real companies routinely sell into more than one category, and a brand scored as deviated may be correctly placed in a category we did not designate as true. The magnet-category concentration is consistent with genuine misfiling rather than labelling noise, and only 10 of 194 deviations share a content word with the true label, but the concern is not fully eliminated.
- **The closed-model comparison is not like for like.** Different sampling depth, an explicit abstention option and visible candidates all favour the closed models. It supports the claim that abstention policy differs, not that the closed models know more.
- **Dimension frames are not mutually comparable.** Each sentiment dimension uses its own prompt frame, so only within-dimension comparisons are reported.
- **Surface coverage is uneven.** Per-variable n ranges from 855 to 998 because not every brand has every surface, and eight censored surfaces are excluded from the raw counts.
- **Cohort is B2B SaaS and English-language.** Nothing here should be extrapolated to consumer brands or to non-English corpora.

7.3 Future work

The obvious next step is a second open-weight family, to separate results that are properties of language models from results that are properties of Gemma. A prospective design would also be feasible for the owned-properties effect: help centres can be published on a schedule, so their arrival could be dated against successive model releases rather than measured cross-sectionally. Finally, the margin structure invites a calibration study, since a reader-visible confidence signal would convert the steepest finding in this paper into something actionable at the point of use.

8 Conclusion

Taking the reading from the weights rather than from generated text separates two failure modes that behavioural measurement conflates. Half of a 1,000-brand B2B SaaS cohort is never named unprompted in its own category on Gemma 4 12B, and one in five of the brands the model will place is placed wrongly, with the model almost never abstaining and never signalling how firmly it holds the answer. The two outcomes have largely disjoint predictors, technical-surface coverage buying presence and a brand's own reference material buying correctness, which means the first question for any brand is which of the two it is failing, not which tactic to run.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

A Full adjusted-effects table

Table 6 lists every variable that survives Benjamini–Hochberg correction within its family on either outcome. Variables are named as they appear in the analysis output: `explains_<host>` is the share of snippets from that host which state the brand's product category, `count_<host>` is raw mention volume on it, and `followers_<host>` is audience size.

B Reproduction

Probe construction. Distractor categories are resampled from the cohort's 333 categories under a word-count constraint with character length as a tiebreak, seeded at 20260828. Sequences are right-padded and scored by mean token log-probability masked by the attention mask, which matches unpadded scoring exactly.

Table 6: All variables surviving Benjamini–Hochberg correction within family, on either outcome. Average marginal effect in outcome points per standard deviation, adjusted for crawl-graph rank and referring domains (and, on accuracy, for the true label's rank position in the probe).

Variable	Family	n	pts/SD	95% CI	p_{adj}
Being named (recall)					
category_crowding	namespace	885	-8.67	[-11.93, -5.42]	<0.001
explains_stackoverflow.com	coverage_language	657	8.17	[3.83, 12.51]	0.004
docs_passages	own_property	560	-6.17	[-10.16, -2.18]	0.039
explains_dev.to	coverage_language	590	5.96	[2.01, 9.90]	0.028
name_is_word	namespace	885	-3.99	[-7.20, -0.79]	0.037
Being described correctly (category accuracy)					
name_is_word	namespace	869	-9.42	[-11.47, -7.38]	<0.001
wiki_article_exists	own_property	869	-8.68	[-11.61, -5.75]	<0.001
wiki_article_length	own_property	869	-8.03	[-10.83, -5.23]	<0.001
mean_snippet_words	coverage_volume	869	-6.03	[-8.72, -3.33]	<0.001
title_share	coverage_volume	869	-5.82	[-8.44, -3.20]	<0.001
yt_third_party_channels	coverage_volume	847	-5.59	[-7.98, -3.20]	<0.001
explains_techcrunch.com	coverage_language	557	5.06	[1.73, 8.39]	0.026
explains_github.com	coverage_language	413	-4.91	[-7.90, -1.91]	0.024
count_theguardian.com	coverage_volume	860	-4.73	[-7.22, -2.25]	0.002
spell_suggested	namespace	869	4.66	[1.72, 7.59]	0.005
n_integrations	own_property	431	4.58	[1.13, 8.04]	0.030
on_theguardian.com	coverage_volume	860	-4.56	[-7.30, -1.83]	0.010
category_label_chars	namespace	869	4.30	[0.89, 7.71]	0.022
n_platforms	own_property	770	3.92	[1.19, 6.65]	0.020
count_trustpilot.com	coverage_volume	864	-3.86	[-6.53, -1.20]	0.030
followers_youtube.com	coverage_volume	633	-3.84	[-6.45, -1.24]	0.030
li_followers_api	coverage_volume	618	3.84	[1.00, 6.68]	0.048
has_help_centre	own_property	869	3.81	[1.16, 6.46]	0.020

Sensitivity to the two probe corrections. The same 1,000 brands give materially different headlines under the three scoring and distractor combinations we ran, which is why both corrections are reported rather than folded silently into the result:

Run	Cons.	Dev.	Unaddr.
Len-matched, right pad	787	194	19
Orig. probes, right pad	753	225	22
Len-matched, left pad	663	320	17

Only the first is reported in the body. The left-padded run is retained for comparability with the original execution and is not a valid measurement, since under left padding the score depends on the longest option in the batch.

Recall matcher. Word-boundary matching with legal-suffix stripping, domain-form acceptance and case rules. The earlier substring matcher is not usable: it matched short brand names inside ordinary words, which in one case inverted a brand's result entirely.

Noise floor. The reseed run repeats 111 categories under seed 20260827 and gives the 13% per-brand flip rate quoted throughout.

Artifacts. Cohort definitions, per-brand outcomes, per-variable model output and the figure-generation scripts are held in the study repository under `experiments/landmark-study/`, with results as JSON in `results/` and analysis code in `scripts/`.