Same Evidence, Different Decisions: Document Segmentation Bias in Retrieval-Augmented Generation

Jiang Hui **Dichun Deng** **Chenyang Zhou** **Zhenrui Zhang** **Chenghui Wu**
Baidu Inc.

Abstract

Document chunking is widely treated as neutral preprocessing in retrieval-augmented generation, but in content-balanced interventions, re-segmenting the same evidence, without changing its semantic content, can shift fixed-choice LLM decisions and flip them in near-threshold regimes. Across full ratio-sweep experiments on 9 primary queries and 6 LLMs, document-unit structure systematically shifts model choices. We fit a shifted sigmoid whose steepness captures structural sensitivity and whose offset c captures fitted decision asymmetry. The two fitted parameters enable post-hoc flip-point estimation after per-query calibration, not a priori prediction. The shifted-sigmoid fit reaches $R^2 \geq 0.80$ in 38/41 non-degenerate model–query combinations; the English experiments show the same form on 9 queries and 4 LLMs. It persists under noisy real-corpus retrieval when retrieved evidence is stance-balanced ($\Delta P = 0.53$, $p < .001$, on one THUCNews retrieval query). The effect is regime-dependent: near-threshold decisions are vulnerable, while large decision offsets resist manipulation. Chunking and boundary placement can therefore affect downstream fixed-choice RAG decisions, with implications for retrieval system design and adversarial robustness.

1 Introduction

Retrieval-augmented generation (RAG) has become the dominant paradigm for grounding LLM outputs in external knowledge (Lewis et al., 2020; Borgeaud et al., 2022). Yet it is unclear whether re-partitioning the same retrieved evidence into different document units affects which position the LLM adopts.

The issue matters for generative engine optimization (GEO) (Aggarwal et al., 2024) and adversarial RAG attacks (Zou et al., 2025; Greshake et al., 2023). Prior work identifies positional biases (Liu et al., 2024) and susceptibility to irrelevant context (Shi et al., 2023), but those studies vary which content is present or where it appears, not how the same content is divided. In natural RAG queries, content, quality, authority, and position are confounded.

We therefore hold evidence content fixed while varying its document-unit structure and measure the resulting decisions. Our core contributions:

1. **Segmentation can causally affect RAG decisions.** In content-balanced interventions, re-segmenting fixed evidence can flip decisions (Figure 2). Across ratio sweeps, document-unit structure systematically modulates LLM outputs (Figure 1), where document-unit balance and content composition co-vary. This broader sensitivity replicates across formatting variations, diverse text conditions, and 6 LLMs, and is directionally consistent in one human-authored-text probe.

2. **A shifted sigmoid summarizes the response** $P(B) = \sigma(k \cdot (\Delta N - c))$ with two fitted parameters. The structural sensitivity k and decision offset c vary substantially across queries, models, and languages, and support flip-point estimation after per-query calibration. A minimal Bayesian accumulation account (the Independence Illusion Hypothesis, §5) offers one behavioral interpretation, but the fit does not depend on that account.

3. **Diagnostic use and boundary conditions.** The sigmoid supports targeted intervention analysis. Once (k, c) have been calibrated for a query, they estimate the document count needed to flip that query's decision within ≤ 2 documents on the 8 testable queries (mean error $= 1.09$); the parameters must be measured per query rather than predicted from document features. In the content-balanced intervention,

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50364–50386
October 24-29, 2026 ©2026 Association for Computational Linguistics

merging to $K{=}1$ reduces the segmentation effect almost entirely, indicating that boundary placement drives much of the observed shift. The effect is *conditional*, appearing primarily in near-threshold regimes ($|c| \leq 2$) where retrieval surfaces mixed-stance documents, observed under stance-balanced public-corpus retrieval.

Scope. We run full ratio sweeps on six API-accessible LLMs, using ERNIE-4.5-turbo for the primary analyses and five additional models for cross-model validation. We also test nine English counterparts of the Chinese queries on four models and conduct a limited real-corpus evaluation on THUCNews (Guo et al., 2016). The main task is two-sided fixed-choice decision making; whether the same segmentation effect extends to free-form generation, summarization, or expressed uncertainty remains open.

2 Related Work

RAG and contextual biases. RAG (Lewis et al., 2020; Guu et al., 2020) has been widely adopted for knowledge-grounded generation, with variants including fusion-in-decoder (Izacard and Grave, 2021), self-reflective retrieval (Asai et al., 2024), black-box augmentation (Shi et al., 2024), and in-context retrieval (Ram et al., 2023). Surveys and analyses catalog growing challenges in RAG pipelines (Gao et al., 2024; Barnett et al., 2024; Zhu et al., 2025). Liu et al. (2024) study positional bias ("Lost in the Middle"); Shi et al. (2023) show irrelevant context degrades performance; Lu et al. (2022) demonstrate prompt-order sensitivity. Closest to the present setting, Levy et al. (2025) show that answer quality shifts with the number of retrieved documents even when total context length is held fixed, and Jin et al. (2024) report a majority-rule tendency, where a retrieval-augmented model favors whichever side of a conflict more passages support. Both change which content is present, and multi-document aggregation (Izacard and Grave, 2021) takes the units as given; we instead hold the retrieved evidence content fixed and vary only how it is divided into document units — how evidence is partitioned rather than where it appears — and fit a two-parameter model of how that structure shapes decisions.

LLM decision biases and sycophancy. LLMs exhibit systematic biases in evaluation (Wang et al., 2024; Chen et al., 2024), require calibration for reliable few-shot decisions (Zhao et al., 2021), and display sycophantic tendencies toward user-stated positions (Sharma et al., 2024; Wei et al., 2024). Turpin et al. (2023) show that chain-of-thought explanations can be unfaithful to actual reasoning. Nearest to the accumulation account of §5.1, Naphade (2026) reports that paraphrased restatements of one passage can be more persuasive to a RAG system than distinct independent support; it varies independence by rewriting the evidence, whereas we leave the supporting content unchanged and alter only its grouping into document units. Relatedly, duplicating a retrieved item leaves multi-hop QA accuracy essentially unchanged while still shifting grounding and evidence-use traces (Jain and Vedam, 2026), a coarser outcome than the near-threshold fixed-choice decisions measured here. The bias also differs from sycophancy: the decision shifts not under user pressure or prompt framing, but with the document-unit structure of the same evidence.

Adversarial RAG and GEO. PoisonedRAG (Zou et al., 2025) and GASLITE (Ben-Tov and Sharif, 2025) demonstrate adversarial document injection; indirect prompt injection (Greshake et al., 2023; Schulhoff et al., 2023) and retrieval corruption (Xiang et al., 2024; Chaudhari et al., 2024) expose further vulnerabilities. Our sigmoid fit bears on this space: the fitted parameters estimate the document count needed to shift a decision, which is relevant for both attack and defense analysis. GEO (Aggarwal et al., 2024) seeks to optimize documents for generative search; the non-linear decision boundary implied by the sigmoid suggests that optimization is threshold-dependent rather than incremental. Document segmentation, from classical topic segmentation (Hearst, 1997) to modern RAG chunking (Kamradt, 2023; Jimeno Yepes et al., 2024; Wang et al., 2025), becomes a relevant factor; that work chooses how to segment documents before retrieval, whereas we vary the segmentation of already-retrieved evidence.

Causal and interpretability methods. Causal inference in NLP (Feder et al., 2022) provides the methodological foundation; amnesic probing (Elazar et al., 2021) and causal abstraction (Geiger et al., 2021) offer tools for mechanism identification; causal mediation analysis (Vig et al., 2020) traces effects through model com-

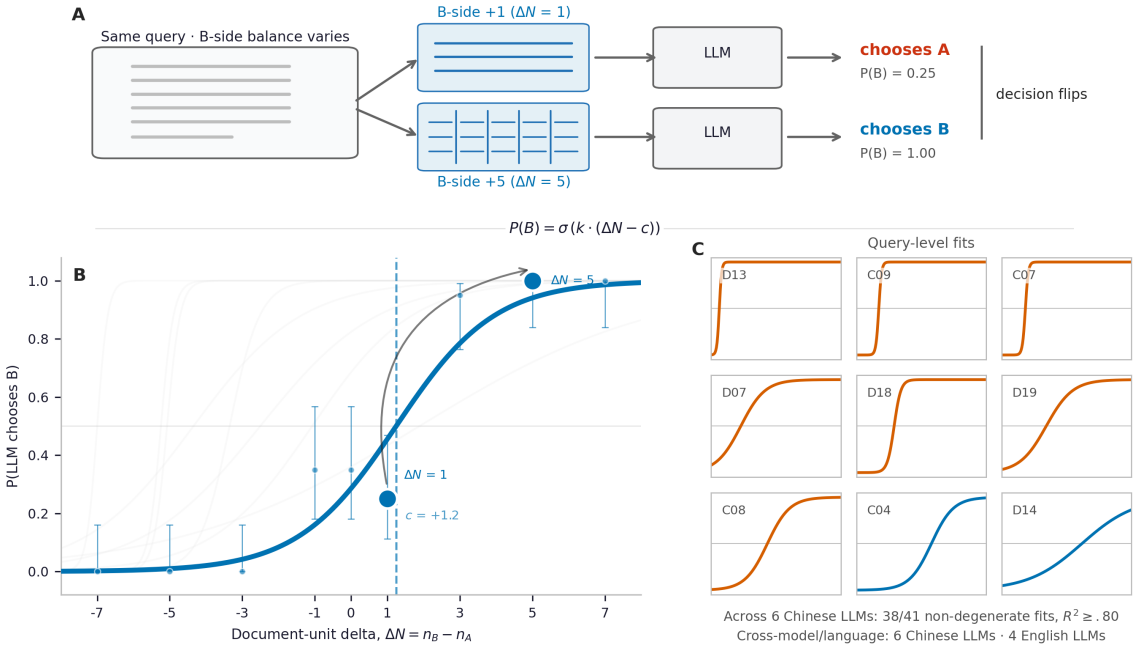


Figure 1: **Document-unit balance shapes LLM decisions. A:** Schematic of the C04 ratio-sweep contrast: evidence drawn from fixed A/B pools is presented with different net B-side document-unit advantages. With a one-unit advantage (ΔN=1), the model mostly chooses A ($P(B)$=0.25); with a five-unit advantage (ΔN=5), it chooses B ($P(B)$=1.00). **B:** For query C04, the highlighted conditions (ΔN=1 and ΔN=5) map to observed points on a shifted sigmoid over document-unit delta $\Delta N = n_B - n_A$. The points cross the fitted threshold c (dashed line); faint gray curves show the other primary queries. **C:** A threshold-shaped pattern appears across the 9 primary queries (blue = count-dominant, $c > +1$; orange = quality-dominant, $c < -1$); 38/41 non-degenerate fits (54 total) achieve $R^2 \geq 0.80$ across 6 Chinese LLMs, and 30/36 across 4 English LLMs (see §4.3).

ponents. We adopt the controlled-intervention paradigm from this literature but target a different question: not *what representations encode* but *how document-level composition shapes behavioral decisions*.

3 Framework

3.1 Task Formulation

We study two-sided decision queries in RAG settings: given query Q and K retrieved documents $\mathcal{D} = \{d_1, \ldots, d_K\}$, the LLM selects one answer from a fixed candidate list. Each list holds four candidates: two *focal* options A and B, which the retrieved documents support in opposite directions, and two non-focal options (typically a compromise between A and B, and an alternative recommendation outside the contrast). Throughout, $P(A)$ and $P(B)$ denote the fractions of *all* valid responses selecting A and B: non-focal selections stay in the denominator, so each measures its option against everything else rather than A against B. We use *document unit* to denote one atomic context unit presented to the LLM, whether it originates from a single retrieved document, a chunk, or a re-segmentation of a larger source. We define *content direction* $\mathrm{dir}(d) \in \{A, B\}$ and the *direction delta* $\Delta N = n_B - n_A$.

Causal intervention design. Following the controlled-intervention paradigm in NLP (Feder et al., 2022), for each of 9 queries, we test 9 document-ratio conditions at deltas $\Delta N \in \{-7, -5, -3, -1, 0, 1, 3, 5, 7\}$, with $n = 10$ trials per condition (shuffled order). A three-layer causal exclusion protocol (authority stripping → data stripping → direction flip) establishes that, within this protocol, content direction alone is sufficient to produce the observed shifts, while authority and data citation are not necessary (quantitative results in §4.1; the five-stage validation pipeline and full protocol in Appendix A).

Unless otherwise noted, fixed-choice experiments use blank-line-separated bracket-labeled document blocks in shuffled order; Appendix E ablates label formats and boundary visibility.

3.2 The Shifted Sigmoid as Parametric Model

We adopt the shifted sigmoid

$$P(B) = \sigma(k \cdot (\Delta N - c)) \quad (1)$$

as our parametric model, a two-parameter extension of simpler baselines (§3.3), and test whether it captures the observed decision landscape. Here k encodes structural sensitivity and c encodes a fitted decision offset (the number of additional B-side documents needed to overcome A's advantage); as a fitted behavioral parameter, c absorbs query-, evidence-, and model-specific factors rather than reflecting an independent quality measure. We refer to fitted regimes with $|c| \leq 2$ as near-threshold on the measured ΔN scale. This functional form is consistent with Bayesian accumulation assumptions (§5.1), but the empirical evaluation is independent of any specific mechanistic account.

3.3 Null Models

We compare the shifted sigmoid against three simpler alternatives to assess whether the fit is non-trivial: Pure Vote (c=0), Flat Sigmoid (global k, no shift), and Per-Query Linear. Both c=0 models fail on all 9 queries (0/9 wins on an AIC-style penalized-fit score), establishing that per-query fitted offset is the non-trivial finding (Eq. 1; Appendix P).

3.4 Cross-Model Validation Design

We replicate core experiments across 6 API-accessible LLMs via commercial APIs: ERNIE-4.5-turbo (primary), Kimi-K2.5, DeepSeek-V3.2, Qwen3-32B, MiniMax-M2.5, and GLM-5.1. Each model is tested with the same prompt template and the same evidence document pools, with document order shuffled independently per trial. For the full ratio sweep, we fit per-model sigmoids on 9 queries $\times$ 9 ratios $\times$ 10 trials (54 model–query combinations). We also validate direction-flip and authority sensitivity on GPT-5, Claude-Sonnet-4.6, Gemini-2.5-Pro, and Qwen3-Max. The model diversity tests whether the phenomenon is specific to a single model family.

4 Results

We first map how choice probability responds to controlled document-unit ratios (§4.1), then test the causal role of segmentation by holding evidence fixed (§4.2), before testing generalization (§4.3) and regime-dependent limits (§4.4).

Query	k	c	R^2	Regime
C04 Coffee	0.74	$+1.24$	0.922	Count-dom
C07 Cooking	5.52	-5.07	0.989†	Quality-dom
C08 Sleep	0.73	-1.07	0.939	Quality-dom
C09 Purifier	5.73	-5.24	0.976	Quality-dom
D07 PC	0.67	-4.34	0.880†	Quality-dom
D13 Math	6.85	-7.00	1.000†	Quality-dom
D14 Exercise	0.30	$+1.89$	0.741	Count-dom
D18 Housing	2.39	-3.35	0.993†	Quality-dom
D19 Flooring	0.62	-2.50	0.893	Quality-dom

Table 1: Shifted sigmoid parameters across 9 primary queries (ERNIE-4.5; generalization in §4.3). † Heavily saturated sweep ($\geq$7 of the 9 ratio conditions exactly at $P(B)$=0 or 1): the slope k is weakly identified, and the step location c is well constrained on D07 and D18 but not on C07 or D13 (D.5). C09 sits just below this cutoff (6 of 9) and should be read the same way. *Regime* labels are shorthand for the sign of the fitted offset (Quality-dom: $c < -1$; Count-dom: $c > +1$), not an independent assessment of evidence quality.

4.1 Core Phenomenon: Content Direction and the Sigmoid Decision Surface

Flipping content direction shifts choice probability by $\Delta P = 0.900$ on average across 6 C-series queries ($n = 5$ per condition). This corresponds to a very large effect. A three-layer causal exclusion ablation on the same queries (Appendix Table 5; $n = 10$ per cell) removes institutional prefixes and then numerical cues, producing small mean changes in $P(A)$ (mean $|\Delta P| = 0.04$ and 0.03 over 12 cells, maximum 0.20), whereas swapping content direction moves it wherever the baseline has headroom (mean $|\Delta P| = 0.19$; 0.55 and 0.40 on the two queries not already at the floor). These results establish that content direction alone is sufficient to produce the effect in this protocol; authority and data citation are not required. The direction-flip contrast reproduces on 5 further models ($\Delta P > 0.8$ on every model–query cell that returned a parseable answer under both conditions; per-model coverage in Appendix A).

Sigmoid fit. The shifted sigmoid (§3.2) achieves $R^2 \geq 0.74$ on all 9 queries (mean = 0.926). Table 1 shows the fitted parameters.

Key findings. The fitted offset c varies substantially across queries. Across the 9 primary queries, c ranges from -7.0 to $+1.9$ (span = 8.9); a 2-

Contrast	Fixed	Manipulated	Alternative
Repeated unit	Source content	Unit count / volume	Information diversity
Content-balanced C04	A/B text and total volume	A-side unit count K	Semantic evidence change
Artifact controls	Evidence contrast	Presentation factors	Non-semantic artifacts

Table 2: Controls isolating segmentation from alternative explanations.

query E2 expansion (C03, C11; Appendix D) extends the observed c range to $+5.25$; auxiliary query sets are reported separately in Appendix G. Seven of 9 primary queries are Quality-dominant ($c < -1$), where one side is favored even when outnumbered in document units; two are Count-dominant ($c > +1$), where the majority side tends to prevail.

The fitted offset is not zero: the two c=0 models achieve 0/9 wins on an AIC-style penalized-fit score (least-squares parameters scored under a binomial likelihood, not likelihood-maximized AIC; Appendix P). A clamped linear model is competitive on 5/9 queries under the same score, so both flexible forms fit adequately; we use the shifted sigmoid as an interpretable (k, c) coordinate system for comparing queries. Some model–query combinations return no identified fit, and one query shows a non-monotonic response (§4.4). The resulting regimes are visualized in (k, c) space (Figure 5, §D.3).

The (k, c) parameterization summarizes each query's response curve with two fitted behavioral parameters. We test it on 9 primary queries and extend it across 6 models and 2 languages in §4.3–4.4.

4.2 Segmentation Causally Shifts Decisions

We isolate segmentation from common alternative explanations through three contrasts (Table 2).

Among these, the content-balanced intervention (Level 2) is the central causal test (Figure 2). The full ratio sweeps of §4.1 supply the fitted (k, c) but vary how much evidence each side contributes, so they are diagnostic rather than causal for segmentation, while the public-corpus retrieval of §4.3 adds naturalistic support with less experimental control. Additional experiments (boundary injection, semantic vs. random splitting, cross-condition comparison) are detailed in Appendix I.

Level 1: Counting without diversity. Replicating a single document N times: sigmoid holds (mean R^2 per text variant $=$ 0.912–0.995 over 4 queries each; per-fit minimum 0.68), indicating that the document-count signal produces

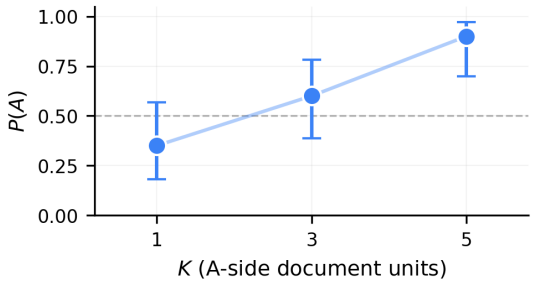


Figure 2: **Segmentation alone flips a content-balanced decision.** For query C04, the A-supporting and B-supporting text is held fixed and only the number of labeled A-side document units varies (K=1: the whole A-side text as a single unit; K=5: the same text divided at sentence boundaries into five units), against three B-side units throughout. Observed $P(A)$ rises from 0.35 to 0.90, crossing the 0.5 decision boundary (dashed). Cohen's $h = 1.23$; error bars show Wilson 95% CI (n=20; two-proportion test $p < .001$). Central content-balanced intervention; large-$|c|$ boundary cases are summarized in §4.4.

shifts even in the absence of information diversity, though total text volume co-varies in this design.

Level 2: Structure change without content change. Holding the A-supporting and B-supporting text fixed and varying only the number of labeled document units the A-side text is presented as (K=1 to K=5, against three B-side units throughout) shifts $P(A)$ from 0.35 to 0.90 on C04 (Cohen's $h = 1.23$). Unlike Level 1, total A-side text volume does not co-vary with K in this design. Segmentation *alone* flips a content-balanced decision (Figure 2; the trial-assembly procedure is given step by step in Appendix A.2). Because the underlying text, its total volume, the prompt template, the answer candidates, and the ordering protocol are all fixed, the 0.35–0.90 shift isolates the effect of presenting the same content as different numbers of document units under randomized ordering. It does not separately identify unit count from the boundary count and unit-position distribution that changing K induces. This manipulation was run on all 17 model–query cells that meet the near-threshold selection criterion, not only on C04: 10 of them cross the 0.5 decision boundary as the A-side unit count rises, spanning four queries and five of the six models. Appendix J reports every cell, including the seven that do not cross and the three checks outside the criterion.

Level 3: Systematic artifact exclusion. A

Model	Query	k	c	R^2
ERNIE-4.5	C04	0.74	+1.24	0.922
Kimi-K2.5	C04	0.92	−1.02	0.952
DeepSeek-V3.2	C04	1.23	−0.92	0.987
MiniMax-M2.5	C04	0.54	−0.41	0.964
Qwen3-32B	C04	1.00	−2.73	0.987
ERNIE-4.5	D18	2.39	−3.35	0.993
Kimi-K2.5	D18	0.84	−2.81	0.985
DeepSeek-V3.2	D18	0.98	−3.19	0.979
MiniMax-M2.5	D18	0.83	−3.45	0.971
Qwen3-32B	D18	6.49	−4.87	0.982

Table 3: Cross-model sigmoid parameters (ERNIE baseline + 4 non-ERNIE models, selected queries; c offsets are model-specific). GLM-5.1 omitted for space; full 6-model table in Appendix F.

four-layer ablation controls for representation format (4/4 formats tested), boundary visibility (visible: 1.0, invisible: 0.5, none: 0.4), position effects (clustered $=$ 0.00 vs. shuffled $=$ 0.70—a position-associated shift), and granularity ($K \in \{1, 2, 5, 10\}$; Table 8, Appendix E).

4.3 Generalization: Cross-Model, Cross-Lingual, and Ecological

We evaluate whether the sigmoid fit generalizes along three axes: across models, across languages, and under stance-balanced real retrieval conditions. For cross-model validation, we run full ratio sweeps on 6 API-accessible LLMs (9 queries $\times$ 9 ratios $\times$ 10 trials per model; 54 model–query combinations, 4,860 total calls). Table 3 shows selected parameters for C04 and D18.

The shifted-sigmoid pattern holds across all 6 tested LLMs ($R^2 \geq 0.80$ in 38 of 41 non-degenerate model–query combinations (54 total; 13 degenerate) across 9 queries; the 13 degenerate combinations returned no identified (k, c), and each has at least 7 of its 9 ratio conditions exactly at $P(B)$=0 or 1). The offset c is model- and evidence-specific, reflecting an emergent property of the query$\times$evidence$\times$model interaction.

Cross-lingual validation. English-language counterparts of the 9 primary Chinese queries (4 C-series and 5 D-series, translated or adapted with matched two-sided answer structure) on 4 API-accessible LLMs (Kimi-K2.5, DeepSeek-V3.2, Qwen3-32B, MiniMax-M2.5): the 30 combinations that produced a converged fit all achieve $R^2 \geq 0.80$ (the remaining 6 did not converge, 4/6 from DeepSeek-V3.2). The same sigmoid form appears in these experiments. EN–ZH c values differ substantially. The offset c is a property of the specific query–document–language interaction rather than an intrinsic query attribute (Appendix L).

Ecological validity. The effect persists under real retrieval when retrieval surfaces balanced evidence, but primarily in a narrower set of balanced contexts.

Controlled retrieval (L1). ERNIE-generated documents retrieved via BM25 (Robertson and Zaragoza, 2009) from a 20-document corpus (5 styles, jieba tokenization, 9 ratio points including densified transition region). Parameters (k, c) from synthetic experiments predict Real-RAG $P(B)$ with $r = 0.75$, MAE $= 0.21$ without re-estimation (Figure 3A). All 4 Real-RAG queries achieve $R^2 > 0.96$ when fitted independently. Separately, on the 8 primary queries whose fitted threshold falls inside the measured ΔN range, the synthetic threshold c predicts flip points within ≤ 2 documents (mean error $= 1.09$; Figure 3B).

Public-corpus retrieval (L2). Documents retrieved via BM25 from a 68K-article subset of THUCNews (Guo et al., 2016) (full corpus: 740K articles; 14 categories), with stance annotation by Qwen3-32B. Of 10 queries, only 3 pass the stance diversity filter ($\geq$3 docs/side in BM25 top-50), and the comparison we report comes from 1 of those 3. This 30% pass rate indicates that real-world BM25 retrieval produces heavily skewed evidence distributions, creating Quality-dominant regimes where one stance naturally dominates.

On the query with genuine diversity (*train vs. plane*), segmentation manipulation produced strong effects: merge-A vs. merge-B yielded $\Delta P = 0.53$ (two-proportion $z = -4.13$, $p < 0.0001$, $n = 30$); split-A vs. merge-A yielded $\Delta P = 0.47$ ($p = 0.0003$). Quality-dominant queries showed $\Delta P = 0.00$, consistent with the sigmoid's prediction that large $|c|$ confers resistance. Negative controls on 7 filtered-out queries: $\Delta P = 0.00$ in 5/7 cases, with two small shifts ($\Delta P = -0.10, -0.125$) at $n \leq 10$.

Additional robustness. With human-authored encyclopedia text, C08 shifts in the predicted direction in both replicates of the balanced configuration ($\Delta P = +0.17$ at n=30 and $+0.27$ at n=15 from K=1 to K=5; pooled $+0.20$, $p = 0.054$), but neither replicate is individually significant and the K=3 point is not ordered at n=15 or pooled, so this check is directionally consis-

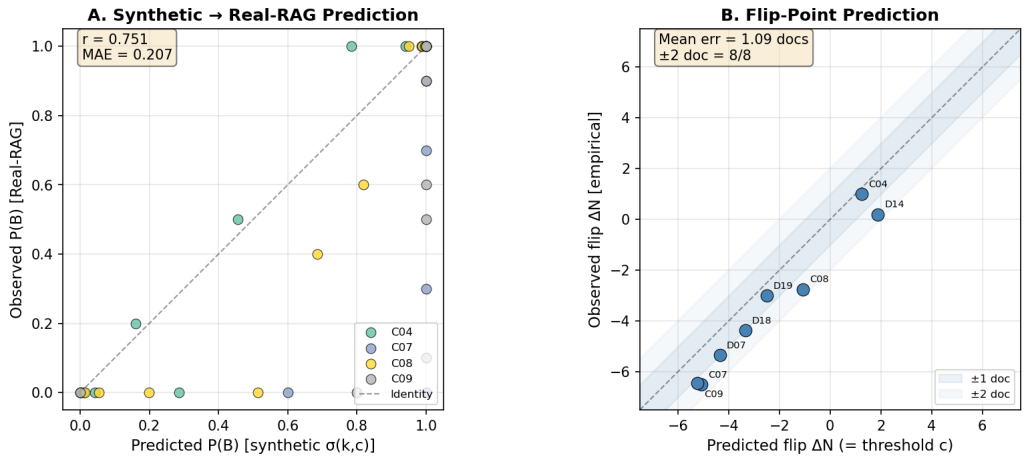


Figure 3: **Predictive validity of synthetic sigmoid parameters. A:** On the 4-query L1 controlled-retrieval set, synthetic sigmoid parameters $\sigma(k \cdot (\Delta N - c))$ predict Real-RAG $P(B)$ without re-estimation ($r = 0.75$, MAE $= 0.21$; each point is one query-by-ratio condition; 36 conditions total). **B:** The synthetic threshold c predicts the empirical decision-flip point (the ΔN at which $P(B)$ crosses 0.5) within ≤ 2 document units for all 8 testable queries (mean error $= 1.09$; D13 excluded: threshold at measurement boundary ($c{=}-7.0 \approx \Delta N_{\min}$)). Shaded bands show ± 1 and ± 2 document-unit tolerance.

tent rather than confirmatory; its companion C04 arm is at ceiling (Appendix I). Temperature and paraphrase checks preserve aggregate effects on the authority and direction conditions they cover, while free-text synthesis remains a boundary condition; none of the three re-runs the segmentation manipulation (Appendix H). Authority framing sensitivity varies $>10\times$ across models ($h = 0.13$–1.70; Appendix K). In the THUCNews evaluation, the dominant limitation is loss of stance diversity rather than a measurable within-query change (Appendix M).

4.4 Boundary Conditions: When Structure Cannot Override Content

Regime-dependent limits. The K-scaling manipulation (Level 2 in §4.2) produces divergent outcomes across regimes. On a Count-dominant query (C04, $c = +1.24$), A-side segmentation from $K{=}1$ to $K{=}5$ flips $P(A)$ from 0.35 to 0.90 ($\Delta P = 0.55$). However, on three queries whose fitted offset places them outside the near-threshold band *for the primary model* (D07, D18, and D19; $c = -4.34, -3.35, -2.50$), the identical manipulation produces $P(A) \leq 0.15$ across all K (0/3 flips). This is consistent with the sigmoid characterization: when $|c|$ is large, $\sigma(k \cdot (\Delta N - c))$ saturates far from 0.5. Segmentation exerts strong effects in near-threshold regimes, but does not override large decision offsets. Band membership is a model–query property rather than a property of the query alone: D19 lies inside the band for four of the other five models and D18 for one of them (Table 10), so these non-responses locate the limit at large fitted offsets rather than at particular topics. In practice, near-threshold regimes arise naturally for contested topics (product comparisons, policy debates) where retrieval surfaces mixed-stance documents.

Instruction defense. Explicit anti-counting instructions produce regime-dependent outcomes: on Count-dominant C04, the sigmoid persists ($R^2 > 0.95$); on Quality-dominant C08, strong instructions collapse it ($R^2 = 0.008$). Instruction-based mitigation is thus incomplete and regime-specific (Appendix N).

Summary. Together, these experiments show a conditional effect: segmentation matters most near the fitted threshold and has little effect at large fitted offsets.

5 Discussion

5.1 A Behavioral Account: Independence Illusion

We use the name **Independence Illusion Hypothesis (IIH)** for one behavioral account of the sigmoid pattern: LLMs behave *as if* each document segment were a semi-independent evidence source, even when segments derive from the same underlying content.

As a minimal model, starting from a prior

$\pi_0 = P(B)/P(A)$, after observing N_A and N_B documents each contributing log-likelihood ratio λ, Bayesian accumulation recovers the observed form: $P(B) = \sigma(\lambda \cdot (\Delta N - c))$, where $c = -\log \pi_0/\lambda$—identifying $k = \lambda$. The logistic transform returns the B share of the A-versus-B contrast, whereas the proportion we plot and fit divides by *all* valid responses (§3.1), so the minimal model approximates the measured quantity rather than equalling it (Appendix F). We do not claim that LLMs literally perform Bayesian updating. The IIH is a behavioral account, not a mechanism: under its Bayesian interpretation, $k = \lambda$ and $c = -\log \pi_0/\lambda$ (Appendix A), but λ and π_0 cannot be measured a priori, so the account is post-hoc rather than predictive.

Post-hoc status. We formulated the IIH *after* observing the sigmoid pattern. Five consistency checks (Appendix B) probe the account: C1–C3 are consistent with the data, C4's token-level probing yields negative rather than the predicted positive log-odds slopes (Appendix C), and C5's open-ended format removes the explicit stance choice while content emphasis still shifts (Cohen's $d = 1.23$).

Token-level caveat. As a diagnostic, we probed first-token log probabilities on Qwen3-8B over 3 queries, varying only the number of A-side document units. The results show partial log-odds linearity but negative slopes on all three: more A-side document units *decrease* $\log P(A)$ at the token level, contrary to the literal Bayesian prediction. Qwen3-8B is not one of the six models whose behavioral fits we report, and its single generated token is also what the behavioral proportion is counted from, so the probe does not compare two levels of one system. We therefore treat IIH as a behavioral abstraction rather than a token-level mechanism; full design and results are in Appendix C.

5.2 Alternative Behavioral Accounts

Several behavioral accounts could partially explain boundary sensitivity: (1) attention redistribution across segment boundaries, (2) redundancy inflation, (3) discourse structure sensitivity, and (4) instruction amplification from labels. Our ablation results (§4.2) show the effect persists with identical documents, so it does not require information diversity; it also appears across label formats, randomized orderings, and invisible boundaries, although label format modulates its magnitude (Appendix E). Individual document features (data citation, consensus cues) produce smaller effects than segmentation ($\Delta P \leq 0.25$; Appendix Q). Overall, the behavioral evidence points to sensitivity to evidence-unit granularity, but the mechanism remains unresolved. Representation-level analysis is needed to distinguish attention redistribution, redundancy inflation, and discourse structure (Geiger et al., 2021; Elazar et al., 2021).

5.3 Practical Implications

The fitted sigmoid enables a lightweight audit workflow: calibrate (k, c) from a ratio sweep, identify near-threshold queries, and stress-test boundary placement. Within a calibrated query the flip point is recovered to ≤ 2 documents on the 8 testable queries (mean error $= 1.09$; Figure 3B), but (k, c) themselves are not predictable a priori (best predictor $\rho = -0.47$, not significant). For RAG, chunking is therefore a decision-affecting design choice rather than neutral preprocessing. An audit can re-run near-threshold queries under alternative chunkings; deduplicating near-duplicate passages, normalizing counts per stance, and aggregating evidence by source rather than by segment may also reduce the effect, but we do not test these or other mitigations. Interactions with embedding, reranking, and hybrid retrieval remain open. The one mitigation we do test, explicit anti-counting instruction, holds in one regime and collapses in another (§4.4, Appendix N).

6 Conclusion

Document boundaries are not neutral formatting: at the behavioral level they act as implicit evidence signals that shape LLM decisions. Segmentation alone, without changing semantic content, can flip content-balanced RAG decisions in near-threshold regimes and a shifted sigmoid with two interpretable fitted parameters describes the behavior.

After calibration a query's flip point is estimable within ≤ 2 documents, useful for auditing but as a diagnostic rather than an a priori predictor.

Predicting c without calibration trials remains an open problem, as does extending the framework beyond two-sided fixed-choice decisions to free-form generation, summarization, or expressed uncertainty.

Limitations

Task and coverage: The primary experiments use Chinese two-sided fixed-choice queries. English validation covers the same 9 domains on 4 LLMs; three C-series queries replace one or both compared options, so this set adds language coverage but no new domains. We do not test whether segmentation affects free-form generation, summarization, or expressed uncertainty, and the effect may differ on topics outside the tested domains. **Evidence and retrieval:** Most evidence documents are synthetic. We partially address this with cross-model replication, one directionally consistent human-authored-text probe on a single query, and stance-balanced real-corpus retrieval (§4.1–4.4); transfer from synthetic to real documents remains uncertain. In THUCNews, 70% of queries fail the stance-diversity filter because BM25 often returns skewed pools (§4.3). The single-corpus result is naturalistic support with less control than the content-balanced intervention, not a second controlled test. **Sampling and model coverage:** Ratio sweeps and segmentation cells use $n = 10$–30 per cell (bootstrap CI for $P(B)$ estimates ≈ 0.10–0.15). Direction-flip, instruction-defense, and paraphrase checks use $n = 5$, and the free-text probe uses $n = 3$, with wider intervals. ERNIE-4.5-turbo is the primary model and five additional LLMs provide cross-model validation. **Parameter stability:** Predicting c from document features alone remains an open problem; the offset captures query×evidence×model interactions that are not reducible to surface features. Fits leave (k, c) unidentified in 13 of the 54 Chinese model–query combinations (24%) and fail to converge in 6 of the 36 English ones (17%); the parameters therefore require per-query calibration rather than a priori prediction (Appendix O). **Interpretation:** The shifted sigmoid describes input–output behavior, not a demonstrated mechanism. We formulated the Independence Illusion Hypothesis after observing the pattern; the consistency checks constrain it but do not establish Bayesian accumulation. The token-level probe uses one small local model (Qwen3-8B), outside the six behavioral models, and finds negative slopes—opposite to the literal Bayesian prediction (§5.1, Appendix C). It reads the same generated token used for the behavioral proportion, so it cannot separate two levels of one system or support a mechanistic claim about the six API models. Our designs also change the number of document units and the number of boundaries together: randomized ordering removes a fixed-position advantage, but unit count is not identified separately from the boundary count and unit-position distribution it induces, so "segmentation" here means that joint manipulation. **Content-balanced flip:** The manipulation crosses the decision boundary on 10 of the 17 near-threshold model–query cells tested, spanning four queries (C04, C08, D14, D19) and five of six models; §4.4 also tests three offset-dominated cases. One crossing is unstable across repeated measurement and is the only crossing on C08. Because the A-side text comes from one source document, the contrast isolates unit count from content change but not from content diversity; the real-corpus setting (§4.3) provides less-controlled evidence on independently authored units.

Ethical Considerations

Our study reveals potential RAG vulnerabilities. We report these findings to enable *defense*, not attack: understanding how document attributes influence LLM decisions is a prerequisite for robust RAG systems. We release controlled experimental stimuli and cached outputs for reproducibility (Appendix R), but not prompt-injection payloads, malicious-document generators, executable attack code, or API-capable attack tools.

Acknowledgments

Large language models provided writing assistance and helped draft and debug analysis code. All experimental designs, analyses, and scientific claims were reviewed and validated by the authors. The authors take full responsibility for all content.

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. 2024. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5–16.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024.

Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*, pages 194–199.

Matan Ben-Tov and Mahmood Sharif. 2025. GASLITEing the retrieval: Exploring vulnerabilities in dense embedding-based search. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 4364–4378.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2206–2240.

Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General backdoor attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.

Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu, and Maosong Sun. 2016. THUCTC: An efficient Chinese text classifier. `https://github.com/thunlp/THUCTC`. Distributed with the THUCNews Chinese news corpus (740K documents, 14 categories).

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880.

Siddharth Jain and Venkat Narayan Vedam. 2026. CUE-R: Beyond the final answer in retrieval-augmented generation. *Preprint*, arXiv:2604.05467.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

Greg Kamradt. 2023. 5 levels of text splitting. GitHub tutorial notebook. Accessed 2026-08-27.

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. 2025. More documents, same length: Isolating the challenge of multiple documents in RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19539–19547, Suzhou, China. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098.

Atharv Naphade. 2026. Rational synthesizers or heuristic followers? analyzing LLMs in RAG-based question-answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40293–40311, San Diego, California, United States. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31210–31227.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 8371–8384.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9440–9450.

Zhitong Wang, Cheng Gao, Chaojun Xiao, Yufei Huang, Shuzheng Si, Kangyang Luo, Yuzhuo Bai, Wenhao Li, Tangjian Duan, Chuancheng Lv, Guoshan Lu, Gang Chen, Fanchao Qi, and Maosong Sun. 2025. Document segmentation matters for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8063–8075, Vienna, Austria. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844.

A Experimental Details

Validation pipeline. The evidence reported in this paper is assembled in five stages.

(i) **Balanced two-sided construction.** Each query fixes a four-candidate list whose A- and B-supporting evidence pools argue in opposite directions, presented as bracket-labeled document units in shuffled order (§3.1; format controls in Appendix E).

(ii) **Content-held-fixed K-scaling.** The A-supporting text is held fixed and only the number of labeled units it is presented as varies, against three B-side units throughout (§4.2, Level 2; trial assembly in §A.2).

(iii) **Shifted-sigmoid fitting.** Ratio sweeps over ΔN are fitted with Eq. 1 per model–query cell and compared against the c=0 and linear nulls (parameter tables in Appendix D, F and L; nulls in Appendix P).

(iv) **Near-threshold selection and follow-up.** Cells whose fitted offset places them near the threshold are selected for the segmentation manipulation and reported cell by cell, crossings and non-crossings alike, alongside the instruction-based mitigation check (Appendix J, N).

(v) **Public-corpus ecological check.** BM25 retrieval over a THUCNews subset with model stance annotation, a stance-diversity filter that most queries fail, and the merge/split manipulation on the queries that survive it (Appendix M).

A.1 Query Definitions

The 9 primary queries, 2 E2 validation queries (C03, C11), and 8 auxiliary expansion queries cover diverse consumer and knowledge domains. Each query defines its four-candidate list (§3.1), its evidence document pools, and keyword lists for automated answer extraction (used only for post-hoc evaluation; not included in model input). Extraction reads the answer field of the model's JSON output and matches it against the candidate list; a response that cannot be resolved to exactly one candidate is counted invalid and dropped from the denominator, whereas a resolved non-focal choice is a valid response that is not A. Full specifications are in the supplementary code.

A.2 Content-Balanced Segmentation Procedure

The content-balanced manipulation (Level 2 of §4.2, plotted in Figure 2 and tabulated for all 17 selected cells in Appendix J) assembles one trial as follows. K is the manipulated-side unit count and $N = 3$ the unit count on the non-manipulated side.

(i) **Fix the supporting content.** The A-supporting and B-supporting text is held fixed for the whole sweep: nothing inside either body of text is added, removed, paraphrased, or reordered at any K, so neither side's text volume changes and the only difference between conditions is how that fixed text is grouped into labeled document units.

(ii) **Group the manipulated side into K document units.** At $K = N$ the A-supporting text keeps its baseline grouping; for $K < N$ the baseline groups are concatenated into K units; for $K > N$ it is concatenated and re-divided into K units at sentence boundaries, with sentences distributed as evenly as the count allows.

(iii) **Hold the other side fixed.** The B-supporting text stays at $N = 3$ labeled units in every condition, and where the query defines one, a single stance-neutral background unit is included as well. The direction delta is therefore $\Delta N = n_B - n_A = 3 - K$, taking the values $+2$, 0, -2 at $K \in \{1, 3, 5\}$.

(iv) **Present and query.** The units are shuffled into a fresh random order for each trial,

rendered as blank-line-separated bracket-labeled blocks (§3), and placed in the fixed-choice template with the query's four candidates at temperature $= 0.3$ (Table 4).

(v) **Score.** Each response is resolved to one candidate by the extraction rule above; unresolvable responses are dropped and $P(A)$ is the share of the remaining valid responses selecting the A-supported candidate, non-focal selections retained in the denominator (§3.1). Steps (i)–(v) repeat n times per condition.

The procedure introduces no new evidence and does not make the units independently authored.

A.3 API Configuration

Models were selected to span diverse model families and providers. Decoding settings are not uniform across experiment families (Table 4): every primary fixed-choice decision trial uses temperature $= 0.3$; Appendix H separately re-runs the authority and direction conditions at $T \in \{0.1, 0.3, 1.0\}$. The other temperatures belong to auxiliary text-generation and judging calls, and top-p is the provider default except in the ERNIE client, which sends top-$p = 0$. The Chinese fixed-choice template is identical across models; the English replication and the log-probability probe use their own templates and answer extractors. The temperature check of Appendix H covers the authority and direction conditions only, not the segmentation manipulation. Specific analyses require model capabilities not universally available (e.g., logprob access for probing); core sigmoid findings are validated across all 6 primary models.

Call family	Model(s)	T	top-p
Fixed-choice decision trials	ERNIE-4.5-turbo	0.3	0
Fixed-choice, cross-model/English	5 + 4 models	0.3	default
Free-text output check	ERNIE-4.5-turbo	0.3	0
Open-ended response generation	ERNIE-4.5-turbo	0.7	default
Stance judgment of free text	Qwen3-32B	0.1	default
Stance labelling of THUCNews docs	Qwen3-32B	0	default
Synthetic evidence generation	model under test	0.7	default
First-token log-probability probe	Qwen3-8B	0.3	default

Table 4: Decoding parameters by call family. "default" means top-p is not sent and the provider default applies. The temperature check in Appendix H re-runs its own conditions at $T \in \{0.1, 0.3, 1.0\}$; all other rows are single-temperature. The log-probability probe decodes one token (`max_tokens` $= 1$).

- **ERNIE-4.5-turbo-128k**: Primary model, 3 QPS, Chinese-native
- **Cross-model validation (5 models)**: Kimi-K2.5, DeepSeek-V3.2, Qwen3-32B, MiniMax-M2.5, GLM-5.1 — via commercial API (full ratio sweeps, §4)
- **Additional models (4)**: GPT-5, Claude-Sonnet-4.6, Gemini-2.5-Pro, Qwen3-Max — via rate-limited commercial API (direction-flip and authority sensitivity, §4.3)
- **Qwen3-8B**: Locally deployed, logprobs enabled (logprob diagnostic; Appendix C)
- Total API calls: ∼32,000+

A.4 Three-Layer Causal Exclusion

We conduct controlled ablations along three document attributes, testing whether removing each factor measurably disrupts the observed effect. Each layer was run on the 6 C-series queries at two document ratios (3A:4B and 5A:2B), $n = 10$ per cell, giving 12 cells per layer; the reported magnitude is the mean absolute change in $P(A)$ against the unablated baseline of the same cell.

Layer	Intervention	Question	Measured shift
1. Authority	Remove institutional prefixes	Is authority necessary?	mean $\lvert\Delta P\rvert = 0.04$ (max 0.20)
2. Data	Remove numerical/statistical cues	Is data persuasiveness necessary?	mean $\lvert\Delta P\rvert = 0.03$ (max 0.10)
3. Direction	Swap A↔B direction	Is content direction sufficient?	mean $\lvert\Delta P\rvert$ = 0.19; 0.55 (C04) and 0.40 (C08)

Table 5: Three-layer causal exclusion: design and measured shift. Layers 1–2 produce only small shifts; layer 3 moves the decision on the two queries whose baseline is not already at the floor. The four remaining queries start at $P(A) \leq 0.10$ and stay there under all three layers, so they neither support nor contradict the direction-flip result.

The same I1 versus I2 direction-flip contrast was run on 5 further models ($n = 5$ requested per cell; some cells returned fewer). Every cell that returned a parseable answer gives $P(A) = 1.0$ under I1 and 0.0 under I2: all 6 queries for DeepSeek-V3.2, GPT-5, Claude-Sonnet-4.6 and Qwen3-Max, and the 4 Gemini-2.5-Pro returned (C09 and C11 timed out). Cells with no parseable answer are dropped rather than scored as $P(A) = 0$. Both conditions sit at the bounds of the scale, so this contrast establishes the direction of movement rather than a graded slope; the ERNIE-4.5-turbo value of 0.900 in §4.1 is the lowest of the six.

A.5 Bayesian Derivation Details

Starting from prior $\pi_0 = P(B)/P(A)$, after N_A A-documents and N_B B-documents, each assumed to contribute an equal log-likelihood ratio λ:

$$\log \frac{P(B \mid \mathcal{D})}{P(A \mid \mathcal{D})} = (N_B - N_A) \cdot \lambda + \log \pi_0 \quad (2)$$

Applying logistic transform: $P(B) = \sigma(\lambda \cdot (\Delta N - c))$ where $c = -\log \pi_0/\lambda$.

Equation 2 is a log-odds between the two focal answers, so the logistic transform of it is the B share of the A-versus-B contrast, i.e. $P(B)/(P(A) + P(B))$. The quantity plotted and fitted throughout the paper instead divides the B count by all valid responses (§3.1), and every query offers four candidates. The two coincide only when no valid response selects a non-focal candidate, which occurs in a small minority of ratio conditions (Appendix F), so the derivation approximates the measured proportion.

Parameter mapping: $k = \lambda$ (per-document evidence weight); $c = -\log \pi_0/\lambda$ (prior–evidence imbalance, the number of additional B-side documents needed to overcome A's content advantage). Both parameters admit interpretable mappings under the Bayesian accumulation assumption, providing a post-hoc interpretation of the fitted sigmoid rather than unconstrained fitting parameters.

B IIH Consistency Checks (Full)

These correspond to checks C1–C5 in §5.1 respectively.

We evaluate five consistency checks related to the IIH, presenting both supporting evidence and boundary conditions where they hold only partially.

1. **C1: k is modulated by content conditions.** All three text conditions (strong, weak, filler) yield sigmoid curves (mean $R^2 = 0.96$ over the 12 fits; minimum 0.68 on one C08 condition), with k varying across conditions. Identical-document sigmoids show $k_{\text{ident}} > k_{\text{diverse}}$ on 3/4 queries (D18 excepted), consistent with document-boundary counting rather than content diversity.

2. **C2: Identical ≈ diverse documents.** Identical-document sigmoids achieve comparable fit quality in the strong-text condition ($R^2 = 0.995$ vs. 0.961 for the diverse-document sweeps of the same four queries), indicating the sigmoid structure persists under reduced content diversity. The 9-query diverse-document mean quoted in §4 is 0.926; the comparison here is restricted to the four queries measured in both designs.

3. **C3: the sweep direction follows the content, not the side.** Swapping the A- and B-side filler documents reverses which content the sweep favors on 2/2 queries: the rank correlation between ΔN and the share choosing the originally B-supported content goes from $+0.872$ to -0.894 on C04 and from $+1.000$ to -0.900 on C08 (five ratio conditions per frame, $n = 10$ trials each). No c is reported for the swapped frame: in content-referenced terms the swapped sweep is decreasing and has no admissible fit under the $k \geq 0$ bound of our fitter (Appendix P), which represents only non-decreasing curves. In the bidirectional quality experiment, graded content strength is associated with dose-responsive shifts in c.

4. **C4: Log-odds linear in K.** Approximate linearity holds on 2/3 queries ($R^2 > 0.78$, $p < 0.05$), but the fitted slope is negative on all three ($\hat{\lambda} = -8.19, -3.51, -10.76$; Table 6), i.e. opposite to the direction the literal IIH predicts. The probe covers a single small locally deployed model that is not among the six whose behavioral fits we report, and it reads the same generated token the behavioral proportion is counted from, so it constrains the literal accumulation account without contrasting levels or models (Appendix C).

5. **C5: Open-ended format removes the explicit stance choice.** In open-ended settings, no explicit stance selection is observed (60/60 responses contain no A/B choice, $n = 30$ per condition), but content emphasis shifts significantly ($p < 0.0001$, Cohen's $d = 1.23$ on A-keyword density), constituting a boundary condition of the fixed-choice setting: the fixed-choice selection effect is not measurable in this format, while content-emphasis sensitivity remains.

Overall, C1–C3 are consistent with the data, C4 holds in form but not in sign on the one model probed, and C5 identifies a boundary condition of the IIH.

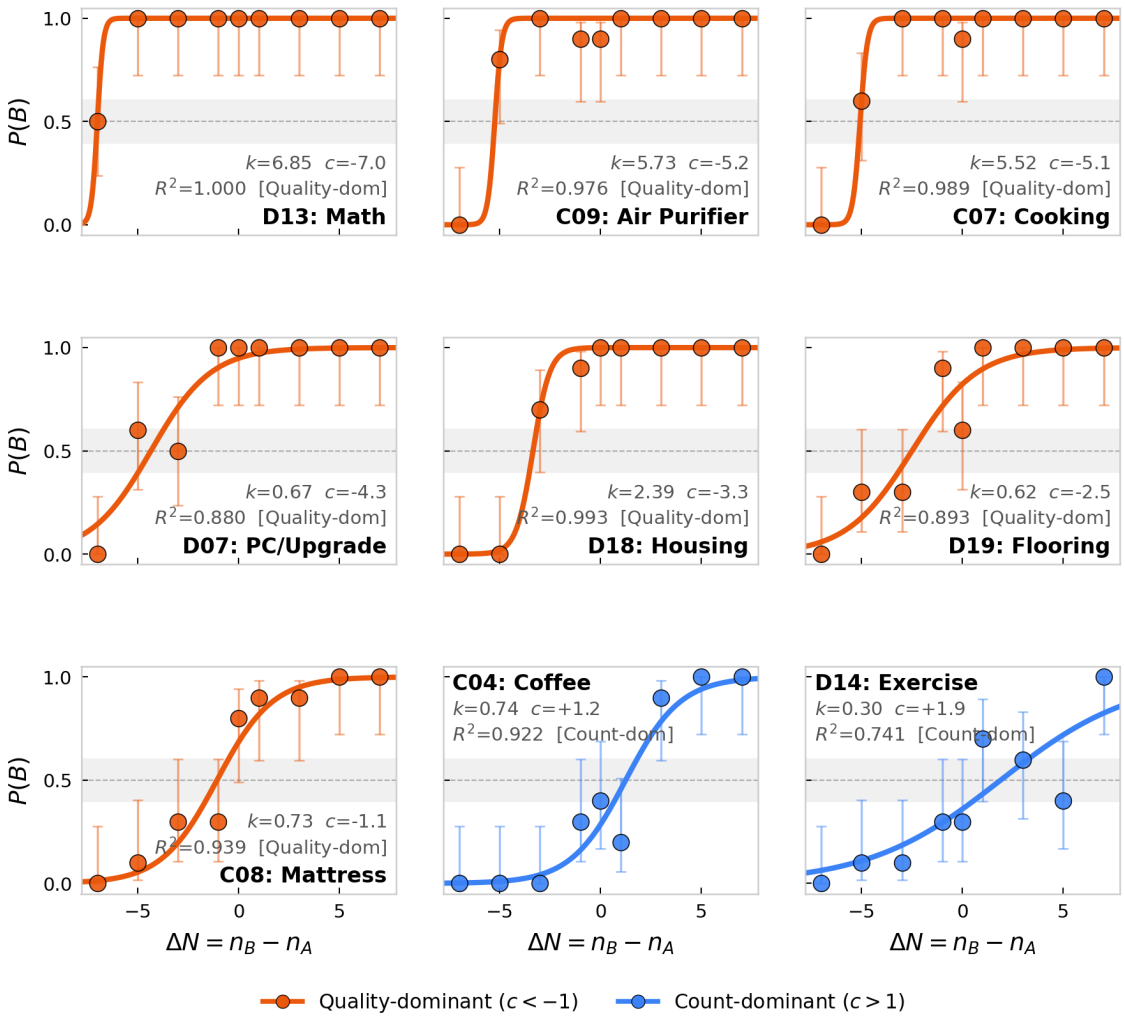


Figure 4: **Shifted sigmoid fits for all 9 primary queries (sorted by c, most Quality-dominant to most Count-dominant).** Circles: observed $P(B)$ with Wilson 95% CI (n=10 trials per ratio, 9 ratio values per query). Smooth curves: fitted $\sigma(k \cdot (\Delta N - c))$ via nonlinear least squares. Heavily saturated panels (C07, D07, D13, D18; $\geq$7 of 9 ratio conditions exactly at 0 or 1) yield high R^2 but poorly constrained k. Parameters match Table 1.

C Logprob Diagnostic Details

Under the literal Bayesian formulation, the IIH implies a testable quantitative pattern: $\log P(A) - \log P(B)$ should increase linearly with the number of A-supporting document units, holding the A-supporting content itself fixed. We probed this on Qwen3-8B (locally deployed, `logprobs=True`, `max_tokens=1`).

Design. For 3 queries (C04 Count-dom, C08 Quality-dom, D18 Quality-dom), the same three A-supporting source documents are divided into $K \in \{1, 2, 3, 5, 7\}$ parts each, giving $3K$ A-side document units against three unsplit B-side documents. The A-side text is identical in every condition, so only the unit count varies and $\Delta N = 3 - 3K$ takes the values $\{0, -3, -6, -12, -18\}$. Division falls on sentence boundaries where a source document contains at least K sentences and on character positions otherwise (7 of 36 document–K combinations divide at sentence boundaries). Log-probabilities of the first output token are extracted; each condition is repeated $n = 10$ times with shuffled order.

Results. The linear form holds on 2/3 queries ($R^2 > 0.78$, $p < 0.05$; Table 6), providing partial support for log-odds linearity. However, the slope is negative on all three: more A-side document units *decrease* rather than increase $\log P(A)$ at the token level, contradicting the literal IIH prediction of positive slope.

Two limits bound this diagnostic: Qwen3-8B is not among the six models whose behavioral fits we report, and with `max_tokens` $= 1$ the probed

log-odds and the generated choice are read from the same output token, so the probe cannot separate a decision from a token-level score within a model.

Query	$\hat{\lambda}$	R^2	p	Regime
C04	-8.19	0.81	0.036	Count-dom
C08	-3.51	0.26	0.377	Quality-dom
D18	-10.76	0.79	0.045	Quality-dom

Table 6: Logprob probing on Qwen3-8B. Linear form holds (2/3 queries), but negative slopes contradict the literal IIH directional prediction. Regression uses the five K-level condition means. $\hat{\lambda}$ is the fitted slope of mean first-token $\log P(A) - \log P(B)$ on K, not on ΔN, and is not an estimate of the behavioral k of Table 1.

D Sigmoid Fitting Details

Figure 4 shows the full sigmoid fits across all 9 primary queries. Each panel displays observed $P(B)$ proportions (Wilson 95% CI) and the fitted sigmoid curve, sorted by c from most Quality-dominant to most Count-dominant.

D.1 11-Query Parameters

Including 2 additional queries (C03, C11) from the E2 sample-boost expansion ($n = 10$ per ratio); both contribute k and c estimates. The primary k–c coupling analysis (D.2) uses the 9 main queries; C03/C11 enter only as a sensitivity check there and in c-predictability analyses (D.4). Table 7 reports all 11 fitted parameters.

Query	Domain	n	k	c	R^2	Regime
C03	Earphones	10	1.75	-5.20	0.991†	Quality-dom‡
C04	Coffee Style	10	0.74	$+1.24$	0.922	Count-dom
C07	Apt Cooking	10	5.52	-5.07	0.989†	Quality-dom
C08	Sleep/Latex	10	0.73	-1.07	0.939	Quality-dom
C09	Air Purifier	10	5.73	-5.24	0.976	Quality-dom
C11	Child Ed	10	8.91	$+5.25$	1.000†	Count-dom‡
D07	PC/Upgrade	10	0.67	-4.34	0.880†	Quality-dom
D13	Math	10	6.85	-7.00	1.000†	Quality-dom
D14	Exercise	10	0.30	$+1.89$	0.741	Count-dom
D18	Housing	10	2.39	-3.35	0.993†	Quality-dom
D19	Flooring	10	0.62	-2.50	0.893	Quality-dom

Table 7: Shifted sigmoid parameters: 11-query set (9 primary + 2 E2-validation). n is the number of trials per ratio condition; every row is a 9-condition sweep at n=10.

†Heavily saturated sweep ($\geq$7 of 9 ratio conditions exactly at $P(B)$=0 or 1): k is poorly constrained; whether c is depends on the sweep (see D.5). C09 is just below the cutoff at 6 of 9. ‡E2-validation query (n=10 per ratio), included only in sensitivity analyses (D.2–D.4), not in primary Table 1.

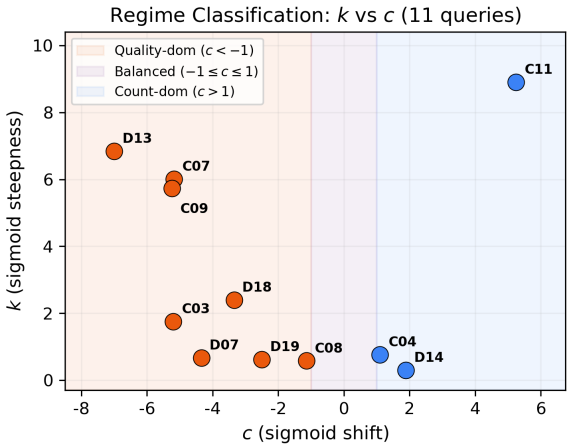


Figure 5: **Regime structure in fitted (k, c) space (11 queries).** Background bands indicate Quality-dominant ($c < -1$), Balanced, and Count-dominant ($c > 1$) regimes. Most queries are Quality-dominant (8/11); Spearman $\rho = -0.37$ (n.s.). C03/C11 are E2-validation queries; heavily saturated fits indicate regimes rather than precise k. The plotted points are the pooled refits described in D.4; recomputing the same correlation on the printed Table 7 values gives $\rho = -0.355$ (D.2). All 11 regime labels are identical under either set.

D.2 k–c Coupling

Correlation is computed on the 9 main queries (including 4 heavily saturated sweeps whose k is weakly identified; see D.5): $\rho(k, c) = -0.800$, $p = 0.010$. Sensitivity check including C03 and C11 from the E2 expansion (same task setup, $n = 10$ per ratio, 11 queries total): $\rho = -0.355$, $p = 0.285$ (both Spearman, on the parameters as printed in Table 7). The shift coincides with the inclusion of C11 ($k = 8.91$, $c = +5.25$; heavily saturated, 8 of 9 conditions at 0 or 1; see D.5), a sign-flipping counterexample to the negative k–c pattern. This analysis is exploratory.

D.3 k–c Regime Scatter

Figure 5 visualizes the regime structure discussed above, showing the separation between Quality-dominant and Count-dominant queries in (k, c) parameter space.

D.4 c Predictability

Predictability remains limited across candidate features. Best predictor: argument diversity ratio ($\rho = -0.473$, $R^2 = 0.229$, $p > 0.05$). Direction accuracy is 72.7% (8/11), suggesting a weak directional tendency. This analysis uses the parameter set stored with it, which differs from Table 7 on

4 of the 11 queries; the largest discrepancies are 0.49 in k and 0.15 in c.

D.5 Failure Modes

These failure modes highlight boundary conditions where the sigmoid abstraction becomes unreliable. D14 ($R^2 = 0.741$): non-monotonic response with wide CIs. C07, D07, D13 and D18 place 7 or more of their 9 ratio conditions exactly at 0 or 1 (C09 is one condition short of the cutoff, at 6 of 9), which yields a high R^2 while leaving k weakly identified: its bootstrap interval spans a factor of 3.5 or more in every one of them. Saturation constrains c unevenly. On D07 and D18 the interval for c stays inside the swept range and spans under three document units, so c is located by where the transition occurs; on C07 and C09 it runs out to the c bound of the fitting box (-15), and on D13 to -7, the most negative ΔN tested, so on those rows the fitted offset identifies a regime rather than a measured threshold. C07 runs the other way under boundary injection: $P(A)$ falls from 0.400 at K=1 to 0.000 at K=5 ($n = 10$ per condition; Spearman $\rho = -0.87$, $p = 0.054$), consistent with quality dominance ($c = -5.07$) overriding the segmentation signal.

E Artifact Ablation

Four-layer ablation design (Table 8):

Layer	Design	Result
L1: Format	4 formats (bracket, source, bullet, plain)	Effect present in all; range varies
L2: Visibility	Visible markers vs. invisible (\n\n) vs. none	C04: 1.0 > 0.5 > 0.4
L3: Position	Clustered vs. shuffled documents	C04: clustered=0.00 vs. shuffled=0.70
L4: Granularity	$K \in \{1, 2, 5, 10\}$ segments	C04 near-monotonic; C08 weak

Table 8: Artifact ablation overview (ERNIE-4.5, C04/C08, n=10 per condition). Values are representative $P(A)$ summaries; full format-level results in Table 9.

L3 reveals that without shuffling, position effects dominate. Randomized ordering removes a fixed-position advantage, but it does not make unit count orthogonal to the boundary count and position distribution that accompany it. L2 shows that even invisible boundaries (double newline) produce partial effects, but explicit labels amplify them.

E.1 Format Effect Summary

Table 9 provides a consolidated comparison of boundary format effects across conditions and models.

Dimension	Condition	ERNIE P(A)	Qwen3 P(A)	Note
Label format (L1, static ratio)				
Source `Source N:`	labeled	1.00	0.80	ERNIE $\Delta P = +0.40$ vs. plain
Bracket `[Doc N]`	labeled	0.90	1.00	ERNIE $\Delta P = +0.30$
Bullet •	labeled	0.90	1.00	ERNIE $\Delta P = +0.30$
Plain (no label)	baseline	0.60	0.10	Qwen3 strongly label-dependent
Boundary visibility (L2)				
Visible (label + newline)	—	1.00	—	Full effect
Invisible (\n\n only)	—	0.50	—	Partial effect
None (space-joined)	—	0.40	—	Minimal
Label format under injection (ΔP = K=5 minus K=1)				
Bullet	injection	+0.70	—	Strongest injection effect
Bracket	injection	+0.60	—	
Newline only	injection	+0.10	—	Weak without explicit labels
Splitting method (ΔP = K=5 minus K=1)				
Semantic (sentence boundary)	—	+0.40	—	Most effective
Random (arbitrary position)	—	+0.20	—	Half effect
Label-only (no actual split)	—	+0.10	—	Small without actual semantic split

Table 9: Format effect summary. P(A) under different boundary formats (C04, 5A:2B, $n = 10$ per cell). ΔP measured relative to plain (no label) condition. Rows summarize separate ablations and are not intended as a single factorial design; dashes indicate conditions not measured in that configuration.

Ranking. Explicit labels $\gg$ semantic split position $>$ label format $\approx$ newline-only. The presence of any document-boundary marker (bracket, bullet, source tag) has a far larger effect ($\Delta P \geq 0.30$) than the specific marker choice ($\Delta P \leq 0.10$ between marker types). Qwen3-32B shows stronger label dependence than ERNIE: without labels, Qwen3 P(A) drops to 0.10 vs. ERNIE's 0.60, suggesting Qwen3 relies more heavily on explicit boundary tokens for document counting.

F Cross-Model Full Table

Table 10 reports fitted sigmoid parameters (k, c, R^2) for all 54 model–query combinations (6 API-accessible LLMs $\times$ 9 queries). The 13 combinations marked "no identified fit" are those on which the fit returned no usable parameters, so k and c are not identified; these are exactly the cells excluded from the 38/41 statistic cited in §4.3. Saturation alone does not force exclusion: 21 of the 54 combinations have $\geq$7 of 9 ratio conditions saturated at 0 or 1, the 13 are a subset of those 21, and the other 8 still yielded an identified fit.

Denominator convention in the sweeps. The plotted and fitted proportions divide by *all* valid responses (§3.1), so a non-focal pick lowers $P(B)$ without raising $P(A)$. Non-focal picks occur in 25 of the 405 Chinese and 23 of the 324 English ratio conditions; the 405 excludes the primary model's 81 conditions, which are stored as proportions only. In two conditions (Qwen3-32B at 4:3 on D19 in Chinese and at 6:1 on D18 in English) *every* valid response was non-focal, so $P(B) = 0$ there records the absence of a B choice

rather than a preference for A; their fits are printed, but neither cell's c is used for a flip-point prediction.

G Expansion Query Set

As an auxiliary scope check beyond the 9 primary queries, we ran the full 9-point ratio sweep on 8 additional D-series queries (distinct consumer-advice topics, not overlapping with the primary set). These are auxiliary, not merged with the primary benchmark, to preserve benchmark integrity (Table 11).

All 8 queries achieve $R^2 > 0.88$. The auxiliary set spans all three regimes (3 Quality-dom, 3 Count-dom, 2 Balanced).

H Robustness

Robustness checks were run on 6 C-series queries with ERNIE-4.5-turbo (Table 12). The four conditions in each check are the authority pair (E1_A, the no-marker baseline, and E1_B, which adds a strong-authority marker to the A-supporting document) and the two direction-flip conditions (E15_I1, E15_I2); none of the three perturbations re-runs the ratio sweep or the K-scaling manipulation. Temperature variation spans $T \in \{0.1, 0.3, 1.0\}$ (n=10 per cell); query paraphrase uses one semantically equivalent rewrite per query (n=5); free-text output uses open-ended responses with post-hoc stance judgment (n=3).

I Boundary Experiments

Table 13 summarizes boundary-related control experiments and key results.

Experiment	Condition	Mean R^2	Key observation
Identical-doc (ERNIE)	Strong	0.995	Counting without diversity
Identical-doc (ERNIE)	Weak+Filler	0.944	k modulated by quality
Identical-doc (Qwen3)	Strong+Filler	0.994	Cross-model replication
Boundary injection	$K = 1, \ldots, 5$	—	3/4 monotonic ($p < 0.05$)
Balanced flip	A-side K=1, 3, 5	—	$\Delta P(A) = 0.55$ on C04

Table 13: Summary of boundary-related control experiments. Mean R^2 is reported for sigmoid-fit sweeps; non-sigmoid rows report directional or probability-shift summaries.

I.1 Semantic vs. Random Boundaries

Comparing sentence-boundary splitting, splitting at random positions (never mid-word), and label-only insertion (2 queries, $n = 10$): on C04 the endpoint effects are ordered (semantic $\Delta P = +0.40$, random $= +0.20$, label-only $= +0.10$), but only the semantic sweep is monotonic in K — the random and label-only sweeps turn at K=3 ($0.1/0.5/0.3$ and $0.3/0.2/0.4$ across $K = 1/3/5$), so the ordering is an endpoint contrast rather than a graded response; on C08 (Quality-dom) only semantic splitting moves $P(A)$ at all ($0.0/0.2/0.4$, against 0.0 at every K in the other two). Boundary effects are not pure label parsing; *where* text splits matters.

I.2 Human-Authored Text

Hand-written encyclopedia-style passages (ERNIE-4.5-turbo, $K \in \{1, 3, 5\}$, A-side content fixed at 5 units against 5 B-side documents) were run twice in this configuration, at n=30 and n=15 per condition. On C08 both replicates move upward between the endpoints—$P(A) = 0.267 \rightarrow 0.300 \rightarrow 0.433$ at n=30 (two-proportion $p = 0.176$) and $0.400 \rightarrow 0.200 \rightarrow 0.667$ at n=15 ($p = 0.143$)—but neither is individually significant, and the interior K=3 point is not ordered in the smaller run or in the pooled data (n=45 per condition: $0.311 \rightarrow 0.267 \rightarrow 0.511$, $z = 1.93$, $p = 0.054$). The C04 arm sits at $P(A) = 1.0$ for every K in both replicates and cannot test the manipulation.

J Follow-up Segmentation Cells: Full Inventory

Table 14 lists every model–query cell on which the content-balanced manipulation of §4.2 was run, including the cells whose outcome does not support the effect. Cells were selected from the 54 fits of Table 10 by the criterion used in §4.4: a non-degenerate fit with $R^2 \geq 0.80$ whose offset lies in the near-threshold band, $|c| \leq 2$, with the non-manipulated side held at three document units. The selection rule was fixed before collecting follow-up outcomes and applied to all 17 selected cells. Seventeen cells meet it, spanning six queries and six models. The other 37 are excluded as $54 = 17 + 22 + 13 + 2$: 22 identified fits whose offset lies farther out ($|c| \geq 2.19$), 13 that returned no usable (k, c) and so cannot be placed relative to the band at all, and two in-band cells whose fits are poor (ERNIE-4.5 on D14, $c = +1.89$, $R^2 = 0.741$; GLM-5.1 on C08, $c = -1.91$, $R^2 = 0.663$). These cells were measured under the settings of §4.2, with the realised per-condition n given in the table.

Outcomes. Ten of the seventeen cells cross the majority threshold as the A-side unit count rises,

Model	Query	k	c	R^2	Regime
ERNIE-4.5	C04	0.74	$+1.24$	0.922	Count-dom
ERNIE-4.5	C07	5.52	-5.07	0.989	Quality-dom
ERNIE-4.5	C08	0.73	-1.07	0.939	Quality-dom
ERNIE-4.5	C09	5.73	-5.24	0.976	Quality-dom
ERNIE-4.5	D07	0.67	-4.34	0.880	Quality-dom
ERNIE-4.5	D13	6.85	-7.00	1.000	Quality-dom
ERNIE-4.5	D14	0.30	$+1.89$	0.741	Count-dom
ERNIE-4.5	D18	2.39	-3.35	0.993	Quality-dom
ERNIE-4.5	D19	0.62	-2.50	0.893	Quality-dom
Kimi-K2.5	C04	0.92	-1.02	0.952	Quality-dom
Kimi-K2.5	C07		no identified fit		
Kimi-K2.5	C08	1.00	-5.22	0.893	Quality-dom
Kimi-K2.5	C09		no identified fit		
Kimi-K2.5	D07	1.04	-4.77	0.973	Quality-dom
Kimi-K2.5	D13		no identified fit		
Kimi-K2.5	D14	0.57	-4.66	0.851	Quality-dom
Kimi-K2.5	D18	0.84	-2.81	0.985	Quality-dom
Kimi-K2.5	D19	1.59	$+0.76$	0.973	Balanced
DeepSeek-V3.2	C04	1.23	-0.92	0.987	Balanced
DeepSeek-V3.2	C07		no identified fit		
DeepSeek-V3.2	C08	0.77	-1.08	0.903	Quality-dom
DeepSeek-V3.2	C09		no identified fit		
DeepSeek-V3.2	D07	0.63	-4.22	0.919	Quality-dom
DeepSeek-V3.2	D13		no identified fit		
DeepSeek-V3.2	D14	0.84	-0.65	0.897	Balanced
DeepSeek-V3.2	D18	0.98	-3.19	0.979	Quality-dom
DeepSeek-V3.2	D19	0.73	-0.69	0.990	Balanced
MiniMax-M2.5	C04	0.54	-0.41	0.964	Balanced
MiniMax-M2.5	C07		no identified fit		
MiniMax-M2.5	C08	0.24	$+3.16$	0.417	Count-dom
MiniMax-M2.5	C09	12.46	-5.99	0.977	Quality-dom
MiniMax-M2.5	D07	3.74	-5.59	0.826	Quality-dom
MiniMax-M2.5	D13		no identified fit		
MiniMax-M2.5	D14	0.55	-1.15	0.930	Quality-dom
MiniMax-M2.5	D18	0.83	-3.45	0.971	Quality-dom
MiniMax-M2.5	D19	0.44	$+0.53$	0.862	Balanced
Qwen3-32B	C04	1.00	-2.73	0.987	Quality-dom
Qwen3-32B	C07		no identified fit		
Qwen3-32B	C08	0.82	-1.30	0.988	Quality-dom
Qwen3-32B	C09	0.44	$+1.84$	0.863	Count-dom
Qwen3-32B	D07	0.74	-4.60	0.912	Quality-dom
Qwen3-32B	D13		no identified fit		
Qwen3-32B	D14	0.41	$+0.78$	0.805	Balanced
Qwen3-32B	D18	6.49	-4.87	0.982	Quality-dom
Qwen3-32B	D19	1.15	$+2.21$	0.998	Count-dom
GLM-5.1	C04	0.51	-0.22	0.868	Balanced
GLM-5.1	C07	6.24	-4.94	0.990	Quality-dom
GLM-5.1	C08	0.28	-1.91	0.663	Quality-dom
GLM-5.1	C09		no identified fit		
GLM-5.1	D07		no identified fit		
GLM-5.1	D13		no identified fit		
GLM-5.1	D14	0.44	$+2.19$	0.874	Count-dom
GLM-5.1	D18	1.74	-0.80	0.995	Balanced
GLM-5.1	D19	14.75	-0.94	0.970	Balanced

Table 10: Full cross-model sigmoid parameters (6 models $\times$ 9 queries, 54 combinations). Degenerate fits (13 total) are combinations on which the fit returned no identified (k, c); every one of them has $\geq$7 of 9 data points saturated at 0 or 1, although 8 further saturated combinations did yield a fit and are retained. The 13 concentrate on D13 (5/6 models), C07 (4/6), C09 (3/6), and D07 (GLM-5.1 only). Of 41 non-degenerate fits, 38 achieve $R^2 \geq 0.80$. Regime labels follow the sign of the fitted c and are not threshold estimates; on the three retained fits with $R^2 < 0.80$ (weakest: MiniMax-M2.5/C08 at 0.417) the label should be read as a sign classification, not as reliable evidence about where that cell's decision threshold lies.

Query	Domain	k	c	R^2	Regime
D01	Remote Work	0.54	-4.25	0.893	Quality-dom
D06	Note Tools	0.79	$+0.05$	0.927	Balanced
D12	Grad School	0.91	-4.27	0.985	Quality-dom
D15	Back Pain	2.14	$+5.99$	0.944	Count-dom
D22	Breakfast	0.91	$+3.94$	0.985	Count-dom
D25	Meal Prep	1.04	$+0.86$	0.969	Balanced
D27	Travel Stay	0.49	-1.89	0.884	Quality-dom
D32	Mortgage	0.49	$+1.40$	0.926	Count-dom

Table 11: Auxiliary D-series expansion (8 queries; ERNIE-4.5, 9 ratio conditions, n=10 per ratio). Not merged with primary benchmark.

Perturbation	Scope	Result	Verdict
Temperature (T=0.1/0.3/1.0)	6q $\times$ 4 cond	Macro spread ≤ 0.084	Stable
Query paraphrase (1 rewrite)	6q $\times$ 4 cond	6/6 preserve I1>I2 order	Stable
Free-text synthesis	6q $\times$ 4 cond	3/4 condition-level agreement	Boundary

Table 12: Robustness and boundary checks (6 C-series queries, ERNIE-4.5-turbo). Temperature reports maximum spread of macro-averaged $P(A)$ across temperatures; paraphrase and free-text checks report qualitative direction agreement. The paraphrase row counts the flip-pair order $P(A)_{\text{I1}} > P(A)_{\text{I2}}$, which holds on 6/6 queries; the stricter requirement that all four conditions stay on the same side of $P(A)$=0.5 as the unperturbed run holds on 3/6 (C03, C08 and C11 each cross it in one condition).

on four of the six queries (C04, C08, D14, D19; on C08 only A1, whose crossing is unstable across repeated measurement) and on five of the six models. Of the seven that do not: two never leave the floor (A2 and A3, both Qwen3-32B), one starts above the threshold and so has no upward headroom (B3), three move in the predicted direction without reaching a majority (A6, G2 and L3), and one reaches exactly 0.500 at K=3 before falling back to 0.100 (L4, discussed below). Because $P(A)$ is the share of *all* valid responses (§3.1), a low $P(A)$ need not mean that B wins: non-focal responses account for $73/90$ in A3, $64/90$ in L3 and $45/90$ in A2. Only in L3 does this explain the non-crossing on its own (A leads B at all three K); in A2 and A3 B still outpolls A (44 versus 1 and 17 versus 0 of 90 valid responses).

Two labels per cell. The endpoint and adjacent-pair rules disagree only on L4: its $P(A) = 0.500$ at K=3 does not persist to either endpoint, so we report it as a non-crossing.

Stability across measurements. Three cells were measured four times each under the same

protocol (A1, A4, B2). A4 and B2 cross in all four. A1 does not: three measurements start at 0.367 and cross, the fourth starts at 0.533 and therefore cannot. At K=1, the condition that decides this, the four runs are statistically indistinguishable (Fisher–Freeman–Halton, $p = 0.50$). Across the nine repeated conditions the largest between-run range is 0.233. A1 is therefore marked as threshold-fragile rather than a stable crossing.

Configuration note. On nine rows the follow-up measurement was made through a different commercial gateway from the one that produced the fit used for selection; the model version requested is the same and no fit was recomputed. All five GLM-5.1 rows use a corrected decoding configuration: earlier runs of B4, O1 and O2 passed a reasoning-disable flag that the endpoint ignored, leaving a biased subsample of parseable responses, and those runs are not printed here.

K Authority Sensitivity

Table 15 reports observed authority effects (Cohen's h) across 6 LLMs under the E1_B vs. E1_A contrast.

Interpretation. These observed effect sizes combine authority framing, model-specific priors, and baseline headroom; this design captures authority effects under realistic RAG conditions rather than isolating them experimentally. We therefore interpret cross-model differences as how authority signals *manifest under model-specific conditions*, rather than as isolated estimates of intrinsic sensitivity, though the variation is *consistent with* training-based accounts.

Outlier pattern. Gemini-2.5-Pro ($h = 1.53$, *) and GPT-5 ($h = 1.70$) both show Large observed effects. Hypotheses consistent with this pattern: (H1) RLHF-style alignment objectives emphasizing deference to external sources (Ouyang et al., 2022); (H2) pretraining data distribution differences (e.g., language-domain effects on Chinese consumer-topic familiarity) leaving more E1_A headroom for authority to add; (H3) model scale (partially disconfirmed; Qwen3-Max is comparable in size yet shows 12× lower h in its one complete run, though two partial runs give $h = 0.29$ and 0.68); (H4) instruction-following fidelity. The two-outlier pattern (both Gemini and GPT-5) is consistent with H1/H2 but further experiments would be required to identify the causal mechanism. The authority cue here is the branded variant alone ("RTINGS [professional review shows]"); the brand-free variant was run only in the pilot's authority-ladder runs, so brand recognition is not separated from source attribution.

L Cross-Lingual Validation

English-language queries covering the same 9 decision domains as the primary Chinese set (4 C-series and 5 D-series), tested on 4 API-accessible LLMs: Kimi-K2.5, DeepSeek-V3.2, Qwen3-32B, MiniMax-M2.5 (36 model–query combinations, 9 ratio conditions per combination, $n = 10$ per condition). Each English query keeps the identifier and four-candidate structure of its Chinese counterpart: the five D-series items are close translations, while the four C-series items keep the decision domain, with three of them substituting one or both compared options. Of the 36 combinations, 6 do not converge on their near-step response patterns and all 30 converged fits achieve $R^2 \geq 0.80$; as in the Chinese sweeps, saturation limits interpretation of k, with 17 of the 36 English sweeps placing at least 7 of their 9 ratio conditions exactly at $P(B)$=0 or 1 (11 of these converged and are printed in Table 16), so we do not read their fitted k as an identified slope (see D.5). The sigmoid *form* is supported cross-lingually in these fixed-choice settings. Across the 22 cells with a converged fit in both languages, the median $|\Delta c|$ is 1.5 document units and the sign of c changes in 8 of the 22, so the fitted offset is not a property of the query alone; the language effect is not separable from the item rewrites. Table 16 reports the full parameter breakdown.

M Ecological Validation Details

M.1 THUCNews Corpus

68K articles, 14 categories, BM25+jieba retrieval, top-50. Stance annotation by Qwen3-32B. 10 queries tested; 3 pass stance diversity filter (≥3/side).

M.2 Noise Degradation

The public-corpus intervention presents the model with only the two focal options, not the four-candidate list of §3.1, so $P(A)$ in this setting is an A-versus-B share rather than the all-responses share used elsewhere in the paper.

Cell	Query	Model	$P(A)$ at A-side K			n	Runs	Majority crossing	AR tier	Stage
			K=1	K=3	K=5					
A1	C08	DeepSeek-V3.2	0.367	0.867	0.733	30	4	yes ↑‡	flip	run
A2	C08	Qwen3-32B	0.000	0.000	0.033	30	1	no (floor)	weak	run
A3	C09	Qwen3-32B	0.000	0.000	0.000	30	1	no (floor)	weak	run
A4	D14	DeepSeek-V3.2	0.167	0.700	0.667	30	4	yes ↑	flip	run
A5	D14	MiniMax-M2.5	0.033	0.333	0.767	30	1	yes ↑	flip	run
A6	D14	Qwen3-32B	0.000	0.400	0.367	30	1	no	strong	run
B1	C04	Kimi-K2.5	0.167	0.400	0.900	30	1	yes ↑	flip	run
B2	C04	DeepSeek-V3.2	0.433	0.767	0.900	30	4	yes ↑	flip	run
B3	C04	MiniMax-M2.5	0.533	0.867	0.800	30	1	no (starts ≥ 0.5)	strong	run
B4	C04	GLM-5.1	0.133	0.433	0.600	30	1	yes ↑	flip	run
G1	C04	ERNIE-4.5-turbo	0.350	0.600	0.900	20	1	yes ↑	flip	prior
G2	C08	ERNIE-4.5-turbo	0.100	0.450	0.250	20	1	no	weak	prior
L1	D19	Kimi-K2.5	0.067	0.567	0.767	30	1	yes ↑	flip	camera
L2	D19	DeepSeek-V3.2	0.133	0.500	0.533	30	1	yes ↑	flip	camera
L3	D19	MiniMax-M2.5	0.200	0.233	0.267	30	1	no	weak	camera
L4	D18	GLM-5.1	0.000	0.500	0.100	30	1	no	flip†	camera
L5	D19	GLM-5.1	0.200	0.367	0.667	30	1	yes ↑	flip	camera
B5	C04	Qwen3-32B	0.667	0.767	0.933	30	1	no (starts ≥ 0.5)	strong	check
O1	D14	GLM-5.1	0.000	0.267	0.200	30	1	no	weak	check
O2	C09	GLM-5.1	0.000	0.000	0.000	30	1	no (floor)	weak	check

Table 14: Every near-threshold cell on which the A-side content-balanced segmentation manipulation was run, plus three additional checks. $P(A)$ is the proportion of valid responses selecting the A-supported answer; n is valid responses per condition. *Runs* is the number of independent measurements of that cell under the same protocol; where there is more than one, the row prints a designated measurement and the supplement lists them all. A further 14 attempts returned no usable sweep (an unavailable model version, wholly failed calls, or a parse collapse); they are present in the released files but are not counted as measurements. *Majority crossing*: strict two-sided endpoint rule; "floor" marks a sweep at $P(A) \leq 0.05$ throughout and "starts ≥ 0.5" one without upward headroom. *AR tier* is the author-response label (adjacent pairs, non-strict comparison, $|\Delta P| \geq 0.25$ = strong); † marks the single disagreement between the two rules and ‡ the single unstable crossing. *Stage*: run = author-response period; prior = pre-existing sweep predating submission; camera = camera-ready stage, August 2026. The three rows below the last rule are checks, not part of the seventeen: B5 and O1 sit outside the selection band ($|c| = 2.73$ and 2.19) and O2's fit is degenerate in Table 10.

train_vs_plane ($n = 30$ valid parses per condition): replacing two of the retrieved stance documents with topically distant articles lowers $P(A)$ from 0.700 under clean split-A retrieval to 0.267 ($\Delta = -0.433$). The dominant degradation is nevertheless the loss of stance diversity: seven of ten queries never reach three documents per side, and the other two that pass the filter record $P(A) = 0.000$ in every clean condition and move by $\Delta P = 0.00$ under the merge/split manipulation (§4.3), a floor with no room for a within-query effect. Under noise injection one of them (*paper_vs_ebook*) does move, from 0.000 to 0.600, so the invariance here is to segmentation, not to retrieval noise.

N Mitigation Resistance

Anti-counting prompts ("focus on argument quality, not document count") were tested on 2 queries $\times$ 3 count ratios (ERNIE-4.5-turbo, $n = 10$ per cell). The prompt moves $P(A)$ in 3 of the 6 conditions and leaves it unchanged in the other 3, and every move is an *increase* in $P(A)$: C04 at 3A:4B, $0.400 \rightarrow 0.800$; C08 at 5A:2B, $0.400 \rightarrow 0.900$; C08 at 4A:3B, $0.300 \rightarrow 0.700$. No sigmoid is fitted for this pilot (three ratio points at $n = 10$), so it supports no claim about c. Prompt-based mitigation is unreliable. This pilot is separate from the full instruction-defense sweep ($3 \times 9 \times 2 \times 5 = 270$ calls) reported in §4.4.

O Model Consistency Analysis

Table 17 catalogs boundary conditions and apparent exceptions, explaining how each relates to the shifted-sigmoid framework.

Model	Obs. h	Level
Qwen3-Max	0.133	Negligible
DeepSeek-V3.2	0.200	Small
ERNIE-4.5-turbo†	0.225	Small
Claude-Sonnet-4.6	0.361	Small
Gemini-2.5-Pro*	1.531	Large
GPT-5	**1.70**	**Large**

Table 15: Observed authority-effect magnitudes (Cohen's h, E1_B vs. E1_A) across the six models in this table, a different set from the six used for sigmoid fitting. h reflects a composite of authority deference, model priors, and baseline headroom; not a pure measure of authority sensitivity.
†ERNIE: 4/6 E1_A baselines at ceiling ($P(A)$=1.0); C08/C09 both $P(A)$=0.6 but only C09 shows authority response (within-query h=1.37).
*Gemini: E1_B from 4/6 queries (C09/C11 timed out); E1_A from 5/6 queries.
The six models were measured across three separate runs; 5 samples per query × condition, macro-averaged across queries. Claude and Gemini were each measured a second time with equal or fuller cell coverage, giving $h = 0.268$ and 1.772 against the printed 0.361 and 1.531; the printed values are therefore run-specific, and that spread is part of the uncertainty in h.

Observation	Context	Interpretation
C04 human-text ceiling	$P(A)$=1.0 at every K	Ceiling leaves no headroom; the manipulation is untestable here, not overridden
C07 boundary reversal	$c = -5.07$	Quality dominance (large $\|c\|$) overrides segmentation signal
E4 behavioral–logit agreement 1/12 (8.3%)	C07/C09 at $P(B)$=1.0	Saturated *behavior*, not the logits, leaves nothing to correspond to; agreement is 7/16 near the boundary
EN–ZH c mismatch	Rewritten items, partly substituted candidates	Fitted offset is not a property of the query alone; language and rewrite effects are not separated

Table 17: Boundary conditions and apparent exceptions under the shifted-sigmoid interpretation.

P Null Model Comparison

Model comparison uses a binomial-likelihood score with an AIC-style parameter penalty: $\ell = \sum_i [n_i^B \log \hat{p}_i + (n_i - n_i^B) \log(1 - \hat{p}_i)]$, score $= 2d - 2\ell$ (where d is the number of free parameters). We write "AIC-style" rather than AIC because the parameters entering ℓ are fitted by unweighted least squares on observed proportions ("Fitting configuration" below) and are then substituted into the binomial likelihood; they are not maximum-likelihood estimates under that likelihood. The score is therefore a coarse penalized-fit comparison, not formal likelihood-based model selection, and we use it only to check that a fitted per-query offset is not a trivial addition. Reported R^2 values are descriptive coefficients of determination ($1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$) computed over condition-level observed proportions and sigmoid-predicted probabilities. Effect sizes use Cohen's h (Cohen, 1988).

Interval methods. Error bars on observed proportions in figures are Wilson score intervals (Wilson, 1927) at $z = 1.96$. The per-condition intervals on $P(B)$ are percentile bootstraps over the trial-level choice vector (2,000 resamples, fixed seed), and intervals on a fitted k or c are percentile bootstraps that resample the ratio *conditions* of a sweep with replacement and refit (500–1,000 resamples); with 9 conditions per sweep these are coarse.

Fitting configuration. The primary fits minimize unweighted squared residuals of observed $P(B)$ against $\Delta N = n_B - n_A$ under box constraints $k \in [0, 15]$ and $c \in [-15, 15]$, with an unweighted R^2; because k is bounded below by zero, a sweep whose tracked share *falls* as ΔN grows has no admissible fit. The cross-model and English sweeps of §4.3 use an unbounded variant with residuals and R^2 weighted by the number of valid responses per condition; the no-usable-estimate cells arise there, where near-step proportions leave k unidentifiable and either drive $|c|$ to implausible magnitudes or prevent convergence. The cross-model and native-evidence scripts also store ΔN as $n_A - n_B$, so their fitted k and c carry the opposite sign to the values tabulated here (the supplement's errata note gives the rule).

- Null-1 (Pure Vote): $d = 0$ (no free parameters)
- Null-2 (Flat Sigmoid): $d = 1$ (one steepness k shared by all queries, fitted at 0.306, with c fixed at 0)
- Null-3 (Per-Query Linear): $d = 2$ (α_q, β_q)
- Shifted Sigmoid: $d = 2$ (k_q, c_q)

Both c=0 null models: 0/9 wins on this score. Shifted sigmoid: 4/9; per-query clamped linear: 5/9 (consistent with §4.1).

Q Feature Decomposition

Table 18 reports exploratory mean $\Delta P(B)$ for four within-document feature toggles (4 C-series queries, ratio 4A:4B, n=10 per condition).

Model	Query	k	c	R^2	Regime
Kimi-K2.5	C04 Coffee	2.26	-1.00	1.000	Balanced
Kimi-K2.5	C07 Cooking	1.07	$+1.32$	0.987	Count-dom
Kimi-K2.5	C08 Sleep	8.89	$+0.25$	0.995	Balanced
Kimi-K2.5	C09 Purifier		no converged fit		
Kimi-K2.5	D07 PC	1.68	-2.24	0.975	Quality-dom
Kimi-K2.5	D13 Math	1.49	-2.03	0.998	Quality-dom
Kimi-K2.5	D14 Exercise	1.86	-1.04	0.996	Quality-dom
Kimi-K2.5	D18 Housing	1.20	-3.77	0.998	Quality-dom
Kimi-K2.5	D19 Flooring	1.37	-0.65	0.999	Balanced
DeepSeek-V3.2	C04 Coffee		no converged fit		
DeepSeek-V3.2	C07 Cooking	1.81	$+3.41$	0.965	Count-dom
DeepSeek-V3.2	C08 Sleep	1.38	$+0.20$	0.993	Balanced
DeepSeek-V3.2	C09 Purifier		no converged fit		
DeepSeek-V3.2	D07 PC		no converged fit		
DeepSeek-V3.2	D13 Math	2.85	-3.48	0.972	Quality-dom
DeepSeek-V3.2	D14 Exercise	2.84	-0.16	0.999	Balanced
DeepSeek-V3.2	D18 Housing	1.44	-4.95	0.995	Quality-dom
DeepSeek-V3.2	D19 Flooring		no converged fit		
Qwen3-32B	C04 Coffee		no converged fit		
Qwen3-32B	C07 Cooking	1.33	$+4.93$	0.962	Count-dom
Qwen3-32B	C08 Sleep	2.21	$+0.25$	0.985	Balanced
Qwen3-32B	C09 Purifier	0.71	-4.57	0.902	Quality-dom
Qwen3-32B	D07 PC	0.68	-2.11	0.828	Quality-dom
Qwen3-32B	D13 Math	1.43	-3.31	0.999	Quality-dom
Qwen3-32B	D14 Exercise	1.02	-0.71	0.964	Balanced
Qwen3-32B	D18 Housing	1.68	-1.76	0.998	Quality-dom
Qwen3-32B	D19 Flooring	1.86	$+1.04$	0.996	Count-dom
MiniMax-M2.5	C04 Coffee	0.61	-2.54	0.969	Quality-dom
MiniMax-M2.5	C07 Cooking	0.35	$+2.16$	0.859	Count-dom
MiniMax-M2.5	C08 Sleep	0.63	-0.24	0.935	Balanced
MiniMax-M2.5	C09 Purifier	1.55	-4.44	0.991	Quality-dom
MiniMax-M2.5	D07 PC	1.04	-2.17	0.988	Quality-dom
MiniMax-M2.5	D13 Math	4.44	-5.49	0.988	Quality-dom
MiniMax-M2.5	D14 Exercise	1.26	-0.23	0.983	Balanced
MiniMax-M2.5	D18 Housing	1.44	-4.95	0.995	Quality-dom
MiniMax-M2.5	D19 Flooring	0.68	-0.16	0.859	Balanced

Table 16: English cross-lingual sigmoid parameters (4 models $\times$ 9 queries, 36 combinations). All 9 English queries are translated or adapted from the 9 primary Chinese queries with matched two-sided answer structure. The 6 combinations without a fit are ones on which curve fitting did not converge (4/6 from DeepSeek-V3.2); all 30 converged fits achieve $R^2 \geq 0.80$. DeepSeek-V3.2 and MiniMax-M2.5 print the same parameters on D18 Housing because their choice counts coincide at all nine ratio conditions: seven are saturated (six at $P(B) = 1$, one at $P(B) = 0$) and the two intermediate conditions return 1/10 and 5/10 A-choices under both models, so the fits are numerically identical; the two cells were nonetheless measured in separate runs.

Feature	$\Delta P(B)$	Interpretation
Data citation	$+0.250$	Positive
Consensus cue	$+0.225$	Positive
Argument structure	$+0.125$	Weak positive
Assertive tone	$+0.100$	Weak positive

Table 18: Exploratory feature-level effects on $P(B)$ (4 C-series queries, n=10 per cell). Values are mean $\Delta P(B)$ when toggling each cue in B-side evidence.

Q.1 Logit Consistency (E4/E4b)

E4 asks whether adding a within-document cue moves Qwen3-8B's first-token log-odds and its behavioral choice in the same direction. For each feature we compare the sign of the change in mean $\log P(B) - \log P(A)$ with the sign of the change in behavioral $P(B)$; a pair also counts as consistent when both changes are negligible ($|\Delta\text{log-odds}| < 0.5$ and $|\Delta P(B)| < 0.05$). At a fixed 3A:4B ratio the two agree in 1 of 12 feature $\times$ query comparisons (8.3%).

The constraint is behavioral, not in the logits: for the two quality-dominant queries in that set (C07, C09) behavioral $P(B)$ is 1.000 in 15 of their 16 conditions, so $\Delta P(B)$ is exactly 0 while the log-odds still move, which the criterion scores as disagreement. Repeating the check at query-specific near-boundary ratios (E4b) raises agreement to 7 of 16 (43.8%, a factor of 5.25); E4b also changes the query set (dropping C07 and C09 and adding C08), so it isolates the effect of moving off the ceiling only up to that substitution.

R Reproducibility Statement

The supplement contains the query definitions, the cached JSON outputs the paper draws on, the figure and inventory code, and a provenance manifest resolving 117 printed results to a source file and JSON path or rule, with eight further entries for sign conventions, per-measurement provenance, cached-file limits, and a manual API call-count estimate with no per-call log. It omits the experiment runners and the fitting and model-comparison code; the recorded per-condition proportions allow the printed fits to be checked and re-fitted with another implementation, but not re-executed.