Supplementary Information for *Cross-Model Memorization Thresholds in Citation Generation: Evidence from Field-Level Cloze on Bibliographic Records*

Anonymous Author

The main text reports paper-cluster bootstrap 95% confidence intervals on per-(condition, field) inflection points and on paired condition–condition contrasts (Table 2 of the main paper and the corresponding subsection). This supplementary file provides a parametric complement: cluster-robust generalized estimating equations (GEE) for both the cross-condition comparison and the within-family reasoning effect. The conclusions track the bootstrap results in direction and significance; the GEE analysis adds the intercept-vs-slope decomposition that the bootstrap (which targets the inflection directly) does not separate.

Specification

For each information-rich field (`title`, `first_author`, `co_authors`, `pages`) we fit a GEE logistic regression of binary recovery on $\log_{10}(\text{cit}) \times C(\text{condition})$ with all seven conditions pooled into a single model, using a binomial family, the logit link, and an exchangeable working correlation structure clustered on the OpenAlex paper identifier (so each paper's seven trials in that field share a working correlation). Sandwich (cluster-robust) standard errors are returned by construction. The omnibus joint Wald test on the $C(\text{condition})$ regressors tests whether the seven baseline recovery levels differ; the omnibus on the $\log_{10}(\text{cit}) \times C(\text{condition})$ regressors tests whether the seven citation–recovery slopes differ.

For the within-family reasoning effect we fit, separately for each of Gemini, GPT-5.4-mini, and DeepSeek, a GEE of the form

$$\text{score} \sim \log_{10}(\text{cit}) \times \text{thinking},$$

again clustered on paper id. The main effect of `thinking` is the intercept shift under reasoning, and the $\log_{10}(\text{cit}) \times$ thinking interaction is the slope shift.

All fits use `statsmodels` defaults (`statsmodels.formula.api.gee`), binomial family, exchangeable correlation, and standard z-tests on the cluster-robust standard errors. The full numeric output is in `data/analysis/gee_results.json` in the project repository, and the fitting script in `scripts/gee_analysis.py`.

Cross-condition omnibus tests (CS-AI)

Per-family reasoning effects, per field (CS-AI)

Table 2 reports, for each model family and each information-rich field, the GEE coefficient on `thinking` (intercept shift) and on $\log_{10}(\text{cit}) \times$ thinking (slope shift), with cluster-robust standard errors and p-values. The qualitative picture matches the per-field bootstrap contrasts in the main paper:

	Condition main effect		$\log_{10}$(cit)× condition	
field	$\chi^2(6)$	p	$\chi^2(6)$	p
`first_author`	36.5	2.2×10^{-6}	8.2	0.22
`title`	50.3	4.1×10^{-9}	20.4	2.4×10^{-3}
`co_authors`	72.0	1.6×10^{-13}	31.8	1.8×10^{-5}
`pages`	172.2	1.5×10^{-34}	93.0	7.1×10^{-18}

Table 1: Joint Wald tests on the seven-condition GEE per field (cluster: `openalex_id`). The condition main effect is significant on every information-rich field. The $\log_{10}$(cit)× condition slope interaction is significant on `title`, `co_authors`, and `pages` but not on `first_author`, meaning that on first author the seven recovery curves differ chiefly in their intercepts, not in their log-linear slopes—consistent with a shared scaling shape and model-specific shifts along the citation axis.

- **Gemini.** No per-field reasoning coefficient—intercept or slope—reaches the conventional $p < 0.05$ threshold. The reasoning condition is statistically indistinguishable from the non-reasoning one on every information-rich field individually, the bootstrap-detected first-author shift of −0.09 decades notwithstanding (the bootstrap pools the within-paper paired structure differently and is somewhat more sensitive to small shifts).
- **GPT-5.4-mini.** The strongest single coefficient is the `co_authors` $\log_{10}$(cit) × thinking interaction ($\beta = +0.52$, $p = 0.002$): reasoning steepens the mini's recovery curve on `co_authors`, which is exactly the field where the non-reasoning mini fares worst. The `title` interaction is marginal ($\beta = +0.20$, $p = 0.089$); on `pages` the main effect is marginal ($\beta = +1.49$, $p = 0.081$). The first-author shift is small and not significant at the field level—the mini's first-author rescue is real (cf. the bootstrap contrast in the main paper) but modest relative to the field's already-small gap.
- **DeepSeek.** The dominant effect is an adverse intercept shift on `title` ($\beta = -0.96$, $p = 1.0 \times 10^{-4}$) and `first_author` ($\beta = -0.60$, $p = 0.045$): reasoning lowers baseline recovery on the identifier fields. On `pages` the main effect is also adverse ($\beta = -0.67$, $p = 0.043$) but the slope steepens ($\beta = +0.42$, $p = 2.7 \times 10^{-4}$), producing the net improvement at high citation counts the main paper notes.

Caveat on apparent disagreements with the bootstrap. A few of the smaller bootstrap-detected effects (e.g. Gemini's −0.09-decade first-author shift) do not reach $p < 0.05$ in the GEE table above. This is expected: the paired-paper bootstrap and the single-coefficient GEE Wald test target different quantities. The bootstrap pairs each iteration's paper resample across the two conditions and looks at $\Delta \log_{10}$(inflection), a single scalar derived from each paired fit; the GEE field test estimates a per-coefficient z on the regression scale within a joint model. For effects that are clearly larger than typical sampling noise (the mini's reasoning rescue, DeepSeek's title/first-author drop) both views agree; for borderline effects the bootstrap is slightly more sensitive because it uses the within-paper pairing directly.

Cross-domain replication: Oncology per-tier accuracy and inflection points

The Oncology master (1,079 papers, `primary_topic.subfield.id = subfields/2730`) was run through the same seven conditions and analyzed with the same pipeline as CS-AI. Tables 3–4 are the direct Oncology counterparts of Tables 1 and 2 of the main paper. Aggregated comparisons

family	field	thinking (intercept shift)		$\log_{10}$(cit) × thinking (slope shift)	
		β (SE)	p	β (SE)	p
Gemini 3 Flash	`first_author`	−0.08 (0.18)	0.65	+0.13 (0.08)	0.12
	`title`	−0.03 (0.19)	0.86	−0.00 (0.08)	0.96
	`co_authors`	+0.04 (0.17)	0.81	+0.09 (0.07)	0.19
	`pages`	+0.02 (0.16)	0.92	+0.08 (0.06)	0.18
GPT-5.4-mini	`first_author`	+0.09 (0.34)	0.80	+0.06 (0.12)	0.60
	`title`	+0.36 (0.32)	0.26	+0.20 (0.12)	0.089
	`co_authors`	+0.47 (0.51)	0.35	+0.52 (0.17)	0.0022
	`pages`	+1.49 (0.85)	0.081	+0.24 (0.27)	0.38
DeepSeek V4-Flash	`first_author`	−0.60 (0.30)	0.045	−0.14 (0.13)	0.25
	`title`	−0.96 (0.25)	1.0×10^{-4}	+0.25 (0.10)	7.9×10^{-3}
	`co_authors`	+0.14 (0.27)	0.59	+0.03 (0.10)	0.79
	`pages`	−0.67 (0.33)	0.043	+0.42 (0.12)	2.7×10^{-4}

Table 2: Per-family reasoning GEE coefficients per field. "Intercept shift" is the main effect of `thinking` (the indicator for the reasoning-enabled condition within a family); "slope shift" is its interaction with $\log_{10}$(cit). Cluster-robust standard errors are in parentheses. Two-sided z-test p-values are shown. For each (family, field) pair the GEE pools the family's two conditions (off/on) into one fit and clusters on `openalex_id`; each paper therefore contributes two trials to the model for that field.

(decade-scale contrasts and the *Cross-domain replication on Oncology* subsection in the main text) are derived from these per-condition values.

	Gemini 3 Flash			GPT-5.4-mini			GPT-5.4 (full)			DeepSeek V4-Flash		
field	low	mid	high	low	mid	high	low	mid	high	low	mid	high
`title`	7	29	72	0	6	39	8	44	82	8	39	82
`first_author`	17	53	88	1	10	50	6	35	75	21	63	87
`co_authors`	1	7	15	0	0	1	0	2	12	2	8	19
`year`	83	81	89	74	78	89	82	80	89	83	80	89
`venue`	78	92	96	36	54	64	54	75	67	72	90	97
`volume`	85	92	99	58	64	74	74	79	92	78	82	91
`issue`	88	97	97	23	21	22	46	55	78	71	70	78
`pages`	31	68	84	2	3	14	15	58	86	19	45	77

Table 3: Per-field cloze recovery accuracy (%) on the Oncology master, for the four non-reasoning conditions, with the six half-decade strata collapsed into three coarse tiers (low: 10–100; mid: 101–1,000; high: > 1,000). Same conventions as Table 1 of the main paper. The field hierarchy (title/first-author memorize first, co-authors and pages later) and the GPT-5.4-mini outlier pattern both replicate; full-tier GPT-5.4 again sits with the other frontier flash-class models. Absolute accuracy levels are uniformly lower than the CS-AI counterparts, particularly on `co_authors` (where no condition exceeds 20% even at the high tier).

Cross-domain replication: Oncology GEE results

The same per-field omnibus tests and per-family reasoning GEEs were fit on the Oncology master (released analysis output in `data/analysis/medicine_gee.json`).

	Gemini 3 Flash		GPT-5.4-mini		GPT-5.4	DeepSeek V4-Flash	
field	off	on	off	on	(full)	off	on
`first_author`	218 [185, 254]	215 [182, 249]	2,519 [1,936, 3,364]	6,312 [4,368, 9,720]	597 [510, 722]	153 [128, 181]	1,719 [1,364, 2,226]
`title`	720 [604, 861]	793 [656, 974]	3,517 [2,766, 4,498]	1,474 [1,256, 1,751]	376 [321, 440]	418 [356, 493]	894 [762, 1,055]
`co_authors`	$4.1\times10^{4\dagger}$ $[1.8, 16]\times10^4$	$3.8\times10^{4\dagger}$ $[1.8, 14]\times10^4$	NA‡ see footnote	$2.2\times10^{4\dagger}$ $[1.3, 4.6]\times10^4$	$2.3\times10^{4\dagger}$ $[1.2, 6.1]\times10^4$	$2.4\times10^{4\dagger}$ $[1.2, 6.5]\times10^4$	$2.2\times10^{4\dagger}$ $[1.1, 5.5]\times10^4$
`pages`	100 [80, 123]	103 [83, 127]	$1.1\times10^{5\dagger}$ $[0.3, 16]\times10^5$	706 [600, 833]	206 [175, 245]	341 [275, 427]	348 [288, 426]

Table 4: Citation count at which the logistic recovery probability crosses 0.5 on Oncology, per condition and field, with 95% paper-cluster bootstrap CIs ($B = 1{,}000$). Same format as Table 2 of the main paper. †Point estimate sits above the 99th percentile of the Oncology citation distribution (≈17,000), so it is an extrapolation of the fitted logistic, not an interpolation; this affects every condition on `co_authors` and GPT-5.4-mini's non-reasoning `pages` fit, because the corresponding curves do not reach $P = 0.5$ within the bulk of the data. ‡The GPT-5.4-mini non-reasoning fit on `co_authors` did not converge on more than 26% of bootstrap resamples (262 of 1,000 iterations finite), so neither point estimate nor CI is reported; the CS-AI counterpart was already extrapolated, and on Oncology the curve is flatter still (cf. Table 3, high-tier accuracy 1%). The overall picture matches the main paper: GPT-5.4-mini is the non-reasoning outlier, the full-tier GPT-5.4 sits with the other frontier flash-class models, and Gemini's reasoning shift is null while DeepSeek's is adverse; the GPT-5.4-mini reasoning rescue helps `title`/`pages` but reverses on `first_author` in this domain.

	Condition main effect		$\log_{10}$(cit)× condition	
field	$\chi^2(6)$	p	$\chi^2(6)$	p
`first_author`	56.8	2.0×10^{-10}	6.9	0.33
`title`	40.1	4.3×10^{-7}	19.3	3.7×10^{-3}
`co_authors`	24.6	4.0×10^{-4}	23.1	7.7×10^{-4}
`pages`	111.5	9.7×10^{-22}	66.9	1.8×10^{-12}

Table 5: Joint Wald tests on the seven-condition GEE per field on Oncology. The condition main effect is significant on every information-rich field, as on CS-AI. The slope interaction follows the same pattern as CS-AI: significant on `title`, `co_authors`, and `pages`, not on `first_author`—the seven first-author curves again differ in intercept only, supporting a single log-linear scaling shared across models with model-specific shifts on this field in both domains.

family	field	thinking (intercept shift)		$\log_{10}$(cit) × thinking (slope shift)	
		β (SE)	p	β (SE)	p
Gemini 3 Flash	`first_author`	−0.39 (0.24)	0.094	+0.18 (0.10)	0.086
	`title`	+0.24 (0.34)	0.48	−0.11 (0.12)	0.35
	`co_authors`	+0.32 (0.43)	0.45	−0.06 (0.13)	0.65
	`pages`	−0.12 (0.13)	0.38	+0.05 (0.06)	0.38
GPT-5.4-mini	`first_author`	−0.09 (0.62)	0.89	−0.21 (0.21)	0.31
	`title`	+0.18 (0.66)	0.78	+0.27 (0.22)	0.22
	`co_authors`	−2.56 (1.54)	0.097	+1.81 (0.59)	2.3×10^{-3}
	`pages`	−1.38 (0.74)	0.063	+1.42 (0.26)	5.4×10^{-8}
DeepSeek V4-Flash	`first_author`	−3.07 (0.52)	3.7×10^{-9}	+0.32 (0.19)	0.097
	`title`	−1.48 (0.44)	7.7×10^{-4}	+0.24 (0.16)	0.14
	`co_authors`	+0.17 (0.52)	0.74	−0.02 (0.16)	0.90
	`pages`	−0.43 (0.25)	0.088	+0.16 (0.10)	0.089

Table 6: Per-family reasoning GEE coefficients on Oncology, in the same format as Table 2. Three patterns differ qualitatively from CS-AI. (i) For GPT-5.4-mini the slope-shift interaction on `pages` is now extremely significant ($\beta = +1.42$, $p = 5 \times 10^{-8}$, against an essentially-null +0.24 on CS-AI), confirming the pages reasoning rescue at very high z; the `co_authors` interaction is also stronger (+1.81 vs. +0.52). (ii) On `first_author` the GPT-5.4-mini slope-shift sign is now negative (−0.21, $p = 0.31$), consistent with the bootstrap-detected reversal of the first-author reasoning rescue noted in the main paper. (iii) DeepSeek's adverse intercept shift on `first_author` is much larger on Oncology ($\beta = -3.07$, $p = 4 \times 10^{-9}$, against −0.60, $p = 0.045$ on CS-AI), and the same shift on `title` is also larger (−1.48 vs. −0.96): in a domain where the identifier is less likely to be memorized, the reasoning-induced displacement of verbatim recall has more room to harm.