

Supplementary Material

The Attribution Crisis in LLM Search Results:

Estimating Ecosystem Exploitation

Strauss, Yang, O'Reilly, Rosenblat, & Moure

Data & Policy

April 2026

A1. Full Hurdle-Model Specification and Diagnostics

The regression is estimated with R's `pscl::hurdle()` function, which maximizes a likelihood that combines a *binary gate* for perfect attribution and a *zero-truncated* negative binomial (NB) for positive gaps. Writing $\pi_i \equiv \Pr(Y_i = 0)$, the log-likelihood is

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha) = \sum_{i: y_i=0} \log \pi_i + \sum_{i: y_i>0} \left\{ \log(1 - \pi_i) + \log f_{\text{NB}}(y_i \mid \lambda_i, \alpha) - \log[1 - f_{\text{NB}}(0 \mid \lambda_i, \alpha)] \right\}, \quad (\text{A1})$$

with $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)^\top$ for the gate and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top$ for the count component. The NB kernel is parameterised by its mean λ_i and over-dispersion $\alpha > 0$:

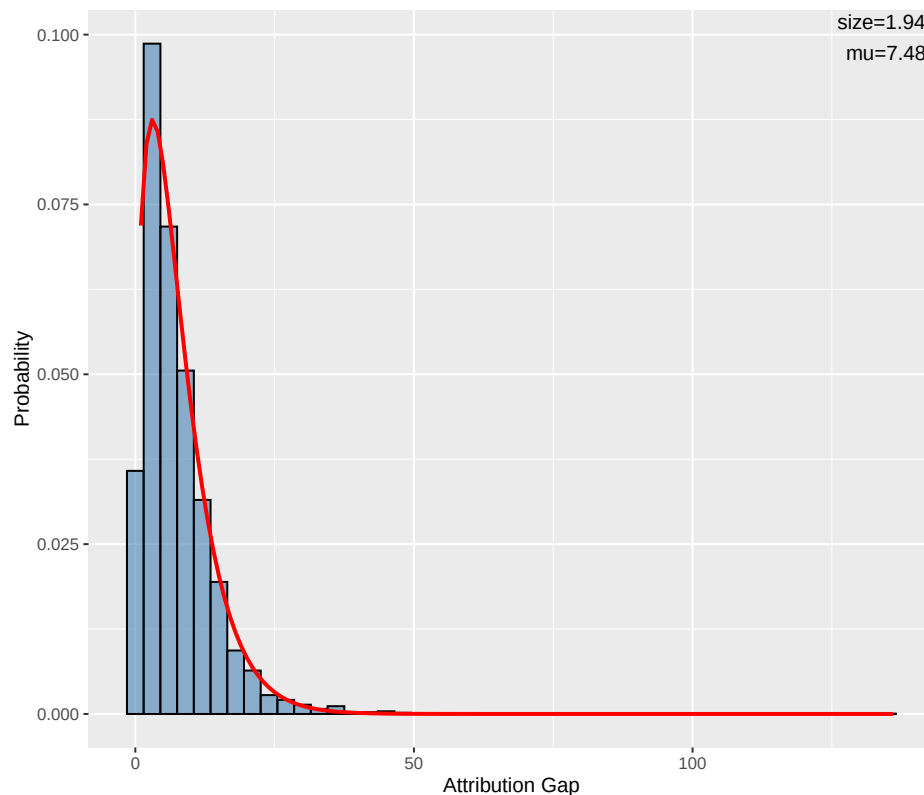
$$f_{\text{NB}}(y \mid \lambda_i, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(\alpha^{-1}) y!} (\alpha \lambda_i)^y (1 + \alpha \lambda_i)^{-(y + \alpha^{-1})}, \quad \text{Var}(Y_i \mid Y_i \geq 1) = \lambda_i + \alpha \lambda_i^2. \quad (\text{A2})$$

Likelihood ratio tests can assess whether the count process is better modeled as Poisson or negative binomial (null hypothesis $H_0 : \alpha = 0$), and whether inclusion of the interaction term improves model fit. A negative binomial outperforms. Model fits (information criteria) are superior with interaction terms and shows domains where the effect of a

model family changes the expected gap magnitude.

With interaction terms the zero-inflated model performs better in a Vuong test than a hurdle model. Our hurdle model, though, aligns better with theoretical understanding of citation process. We believe interpretability advantages outweigh improvements though. BFGS optimization converged in 54 iterations, giving $\hat{\theta} = 5.69$.

Figure A1. Attribution Gap Histogram vs Negative Binomial (Red)

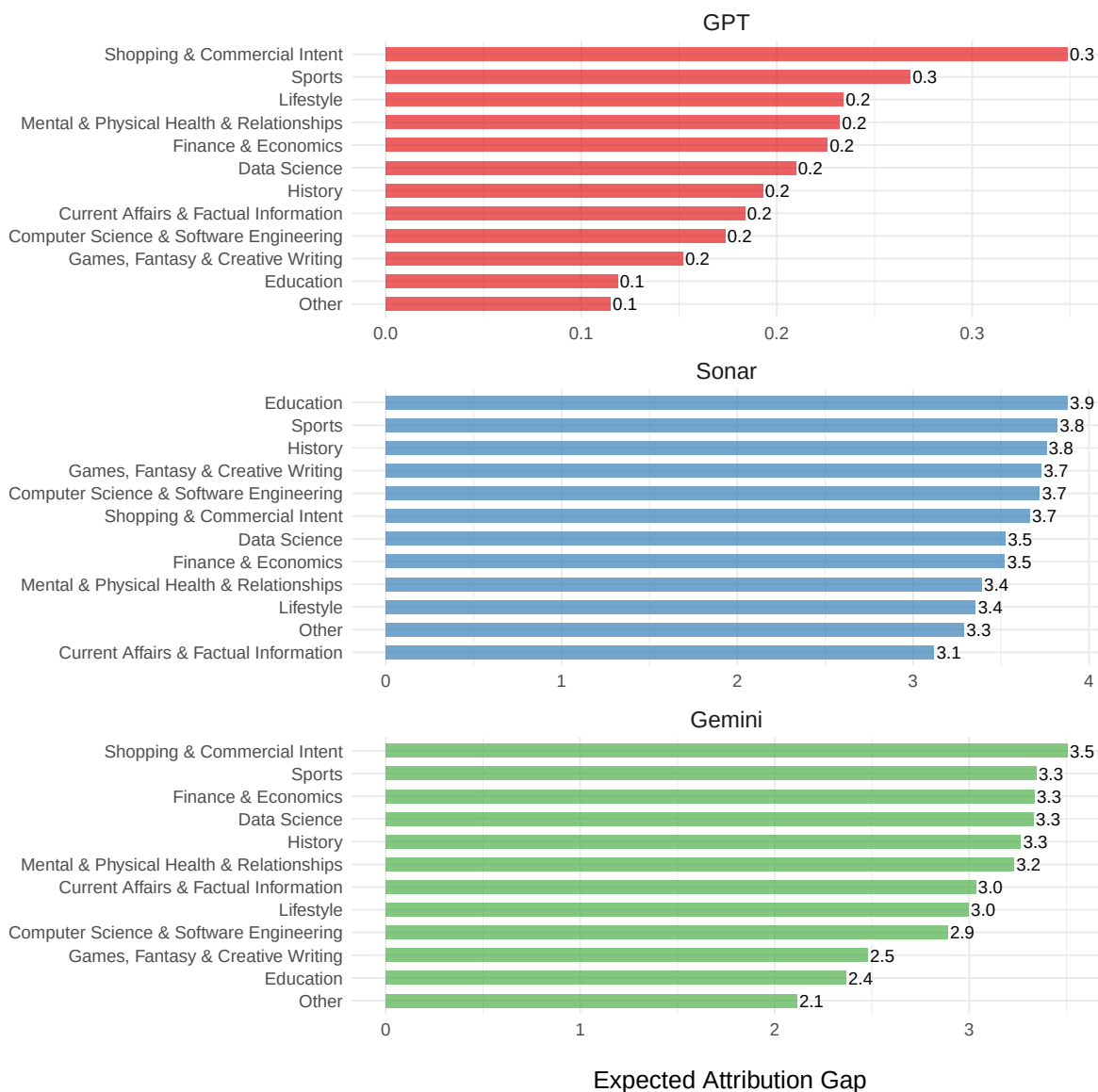


Note: Attribution gap (relevant sites visited in logs minus websites cited by in-text URLs) shows a negative binomial distribution (red over-plot).

A2. Detailed Model Results

Figure A2 shows topic-specific gaps, as predicted by our model. ‘Shopping & Commercial Intent’, perhaps surprisingly, shows the largest gap across models – being 0.2 lower for the Sonar model family. This could reflect a greater selectivity in which results are shown by the model given potential greater commercial incentives to ‘get it right’, or to abide by pre-existing licensing agreements. But it also could simply highlight the absence of a worked-out business model in this area, despite large opportunities to monetize website traffic in this area of LLM search.

Figure A2. Expected Citation Gaps by Regression Model (Query Classification and Model Family)



Note: Bars show the expected citation gap (relevant websites visited in the logs minus websites cited in the LLM's output) for each query classification. Classification labels are ordered by gap magnitude within each panel.

Table A1. Count Model Coefficients (No Interactions)

Term	Estimate	Std. Error	z value	p-value
(Intercept)	2.163	0.078	27.66	0.000
modelfamilyGPT	-1.099	0.156	-7.05	0.000
modelfamilySonar	0.207	0.054	3.81	0.000
classificationCurrent Affairs & Factual Information	-0.021	0.053	-0.39	0.695
classificationData Science	0.057	0.064	0.89	0.371
classificationEducation	-0.115	0.091	-1.26	0.209
classificationFinance & Economics	0.034	0.068	0.50	0.615
classificationGames, Fantasy & Creative Writing	-0.053	0.074	-0.72	0.473
classificationHistory	-0.008	0.087	-0.09	0.929
classificationLifestyle	-0.028	0.064	-0.43	0.665
classificationMental & Physical Health & Relationships	0.050	0.077	0.66	0.512
classificationOther	-0.093	0.065	-1.43	0.154
classificationShopping & Commercial Intent	0.021	0.061	0.35	0.727
classificationSports	-0.033	0.091	-0.37	0.712
search results count	0.125	0.003	39.74	0.000
log(response length)	-0.171	0.008	-20.24	0.000
log(θ)	1.740	0.031	55.39	0.000

Note: Gemini as baseline model. Showing coefficients (main effects) from the count component of the negative binomial hurdle model.

Table A2. Hurdle Model Coefficients (No Interactions)

Term	Estimate	Std. Error	z value	p-value
(Intercept)	-0.729	0.207	-3.52	0.000
modelfamilyGPT	-3.140	0.069	-45.66	0.000
modelfamilySonar	1.601	0.061	26.28	0.000
classificationCurrent Affairs & Factual Information	0.195	0.096	2.03	0.043
classificationData Science	0.268	0.117	2.29	0.022
classificationEducation	-0.245	0.144	-1.69	0.090
classificationFinance & Economics	0.341	0.121	2.83	0.005
classificationGames, Fantasy & Creative Writing	-0.265	0.108	-2.45	0.014
classificationHistory	0.398	0.169	2.36	0.018
classificationLifestyle	0.177	0.111	1.59	0.113
classificationMental & Physical Health & Relationships	0.189	0.144	1.31	0.190
classificationOther	-0.529	0.098	-5.39	0.000
classificationShopping & Commercial Intent	0.567	0.115	4.93	0.000
classificationSports	0.577	0.151	3.81	0.000
log(response length)	0.165	0.024	6.92	0.000

Note: Gemini as baseline model. Showing coefficients (main effects) from the hurdle component of the negative binomial hurdle model.

Table A3. Expected Citation Gaps by Model Family and Topic Classification

Classification	GPT-4o		Gemini		Sonar	
	Est.	95% CI	Est.	95% CI	Est.	95% CI
Computer Science & Software Engineering	0.17	(0.14–0.22)	2.89	(2.63–3.14)	3.72	(3.55–3.89)
Current Affairs & Factual Information	0.18	(0.15–0.23)	3.03	(2.78–3.29)	3.12	(3.00–3.26)
Data Science	0.21	(0.16–0.27)	3.33	(2.97–3.73)	3.52	(3.34–3.74)
Education	0.12	(0.08–0.19)	2.37	(1.97–2.78)	3.88	(3.57–4.19)
Finance & Economics	0.23	(0.18–0.29)	3.34	(2.95–3.73)	3.52	(3.34–3.72)
Games, Fantasy & Creative Writing	0.15	(0.11–0.21)	2.48	(2.18–2.82)	3.73	(3.54–3.92)
History	0.19	(0.13–0.30)	3.27	(2.77–3.80)	3.76	(3.45–4.10)
Lifestyle	0.23	(0.17–0.31)	3.00	(2.69–3.34)	3.36	(3.19–3.54)
Mental & Physical Health & Relationships	0.23	(0.17–0.32)	3.23	(2.76–3.73)	3.39	(3.17–3.63)
Other	0.11	(0.09–0.15)	2.11	(1.88–2.39)	3.29	(3.13–3.47)
Shopping & Commercial Intent	0.35	(0.27–0.45)	3.51	(3.16–3.85)	3.66	(3.48–3.86)
Sports	0.27	(0.19–0.38)	3.35	(2.85–3.95)	3.82	(3.59–4.07)

Note: Total expected attribution gap from negative binomial hurdle model. Predictions are at the median query (5 search results and 2,089 characters). Confidence intervals from parametric bootstrap (n=1,000). Est. = Expected citation gaps per query.

A3. Head-to-Head Regressions: Table of Results

Table A4. Incremental citations per extra URL opened, by model variant

Model variant	Family	n	$\widehat{\beta}_{1m}$	SE	p
ppl-sonar-reasoning-pro-high	Sonar	848	0.447	0.015	1.4×10^{-138}
ppl-sonar-reasoning	Sonar	1635	0.445	0.012	4.3×10^{-215}
gemini-2.0-flash-grounding	Gemini	1189	0.427	0.013	6.0×10^{-172}
api-gpt-4o-search-high-loc	GPT	1222	0.426	0.010	6.1×10^{-240}
api-gpt-4o-search-high	GPT	1695	0.389	0.010	1.3×10^{-224}
ppl-sonar	Sonar	1200	0.339	0.012	1.8×10^{-128}
gemini-2.5-pro-grounding	Gemini	1200	0.312	0.011	4.8×10^{-130}
api-gpt-4o-search	GPT	1188	0.264	0.012	2.5×10^{-89}
api-gpt-4o-mini-search	GPT	1158	0.251	0.012	1.1×10^{-80}
ppl-sonar-pro	Sonar	1208	0.200	0.014	9.5×10^{-46}
ppl-sonar-pro-high	Sonar	1359	0.191	0.012	9.9×10^{-51}

Note: n = pair-wise query comparisons involving the focal model. $\widehat{\beta}_{1m}$ is the coefficient on focal_search_diff, interpreted as the expected number of additional citations produced for each extra URL the focal model opens.