

From Citations to Contributions

LLM-Assisted Credit Scoring of Research Articles



Sana Ebrahimi
University of Illinois Chicago

Suraj Shetiya
IIT Bombay

Abolfazl Asudeh
University of Illinois Chicago



MOTIVATION

A citation counts the same, whatever its role



Broad background

Cited once in the introduction to set the scene.

+1 citation



Building block

A method the paper adapts and extends.

+1 citation



Central idea

The work without which the paper would not exist.

+1 citation

Citation counts, and graph rankings built on them, **ignore what a citation contributes** to the citing paper.

CONTRIBUTION-BASED CREDIT SCORING

How much of a paper's credit is its own, and how much belongs to the work it builds on?

Original contribution $\sigma_{\text{orig}}(d)$

c_1

c_2

c_3

...

A paper's credit totals 1 and divides into its original contribution plus the share each cited work c earns.

Credit allocation as a cooperative game

The game

Players are the paper d , standing for its original contribution, and every reference it cites.

$$N = \{d\} \cup C(d), \quad \nu(N) = 1$$

The Shapley value gives each player its expected marginal contribution.

Why it breaks

It needs a value for every subset: $\nu(S)$, $S \subseteq N$

Each such value is a counterfactual paper that was never written.

Subsets to evaluate: $2^{|C(d)|}$

Instead, we use the structure the paper already has: **its section hierarchy**.

FRAMEWORK

The contribution tree

The root is the whole paper: $\sigma(d) = 1$

Each edge carries a weight, and siblings sum to 1:

$$w(u, v) \in [0, 1], \quad \sum_{v \in \text{ch}(u)} w(u, v) = 1$$

$$\sigma(v) = \prod_{u \in \text{path}(r \rightsquigarrow v)} w(\text{par}(u), u)$$

Hierarchical conservation

Efficiency

Additivity

Shapley-style properties, proved in Theorem 1.

d : **Native Sparse Attention (Yuan et al., 2025)**

w_1 Introduction

w_{1j} **Paragraph with citations**

Own text $\sigma_{\text{orig}}(p)$

b_1 Zhang · Li · Zhou $\sigma(b_1)$

b_2 Tang · Xiao · Gao $\sigma(b_2)$

b_3 Chen · Liu · Desai $\sigma(b_3)$

Each block is scored from its citation context

w_2 Methodology

w_{21} Background

w_{22} Overall Framework

w_3 Experiments

w_4 Efficiency Analysis

w_5 ...

LLMs as noisy comparative estimators

$$X(v | u) = \alpha(u) w(u, v) + \eta_{u,v}$$

Raw scores are comparable only among siblings: each parent has its own scale and noise.

01

Query

Show the parent and its children; ask for 100 points split across them.



02

Normalize

Divide each answer by its sibling sum.



03

Average

Average 5 valid samples; weights still sum to 1.

Requirement: beat the uninformed guess $w_0(u, v) = \frac{1}{|\text{ch}(u)|}$, tested against human annotations.

From citation blocks to cited papers

strategies: KV-cache eviction methods (Zhang et al., 2023b; Li et al., 2024; Zhou et al., 2024), blockwise KV-cache selection methods (Tang et al., 2024; Xiao et al., 2024a; Gao et al., 2024), and sampling, clustering or hashing-based selection methods (Chen et al., 2024b; Liu et al., 2024; Desai et al., 2024). More related sparse attention methods are introduced in Appendix A. Despite these promising strategies, existing sparse attention methods often fall short in practical deployments. Many approaches fail to achieve speedups comparable to their theoretical gains; moreover, most methods lack effective training-time support to fully exploit the sparsity patterns of attention.

 Citation context, sent to the LLM

 b_1
 b_2
 b_3

Paragraph's citation share: $\sigma_{\text{cite}}(p)$

↓ LLM scores each block from its context

 $\sigma(b_1)$
 $\sigma(b_2)$
 $\sigma(b_3)$

↓ ÷ 3

↓ ÷ 3

↓ ÷ 3

Zhang

Li

Zhou

Tang

Xiao

Gao

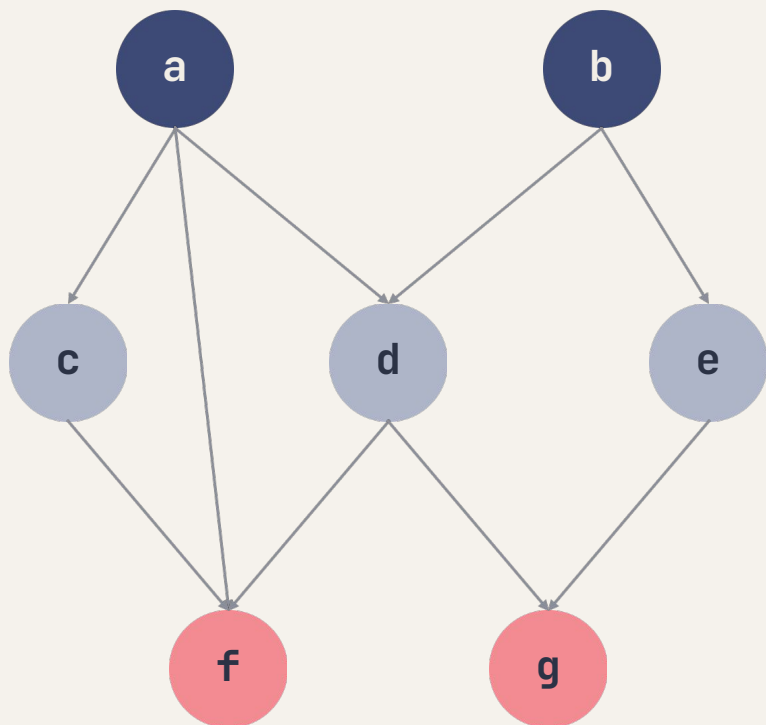
Chen

Liu

Desai

Each cited paper in a block gets an equal share: $\sigma_b(c) = \frac{\sigma(b)}{|C(b)|}$

Credit flows along a weighted citation graph



Edges run from citing to cited paper.

Each citation edge is weighted by the credit the citing paper's tree assigns to the cited one:

$$d \rightarrow q: \sigma_d(q)$$

Every paper starts with 1 unit, keeps its original share, and passes the rest to its references.

$$\sigma_{\mathcal{P}}(d) = \sigma_{\text{orig}}(d) \left(1 + \sum_{\pi \in \{a \rightsquigarrow d\}} \prod_{(u \rightarrow v) \in \pi} \sigma_u(v) \right)$$

$$\sum_{d \in \mathcal{P}} \sigma_{\mathcal{P}}(d) = |\mathcal{P}| \quad \text{Theorem 2}$$

$$O(|\mathcal{P}| + |E|) \quad \text{via topological order}$$

Two scores for every paper

Corpus-level contribution

$$\sigma_{\mathcal{P}}(d)$$

The original credit a paper keeps, plus what it inherits through citation paths. It includes the paper's own unit of self-credit.

Normalized influence

$$\sigma_{\mathcal{P} \setminus d}(d) = \sigma_{\mathcal{P}}(d) - \sigma_{\text{orig}}(d)$$

$$I_{\mathcal{P}}(d) = \frac{\sigma_{\mathcal{P} \setminus d}(d)}{\sum_{d' \in \mathcal{P}} \sigma_{\mathcal{P} \setminus d'}(d')}$$

The share of all cross-paper influence mass that reaches d.

In small or sparse corpora, self-credit dominates the corpus-level score. **Normalized influence isolates the credit a paper receives from others.**

EVALUATION SETUP

50 expert-annotated papers, three references

50

papers annotated by
their authors or domain
experts

3

references: human
experts, Claude Sonnet
4.6, GPT-OSS-120B

8

small local LLMs as
estimators, llama3.2:1b
to phi3:medium

2

non-LLM baselines:
citation and
length-weighted
frequency

Human annotators

Score top-level sections and pick the 4 most foundational citations.

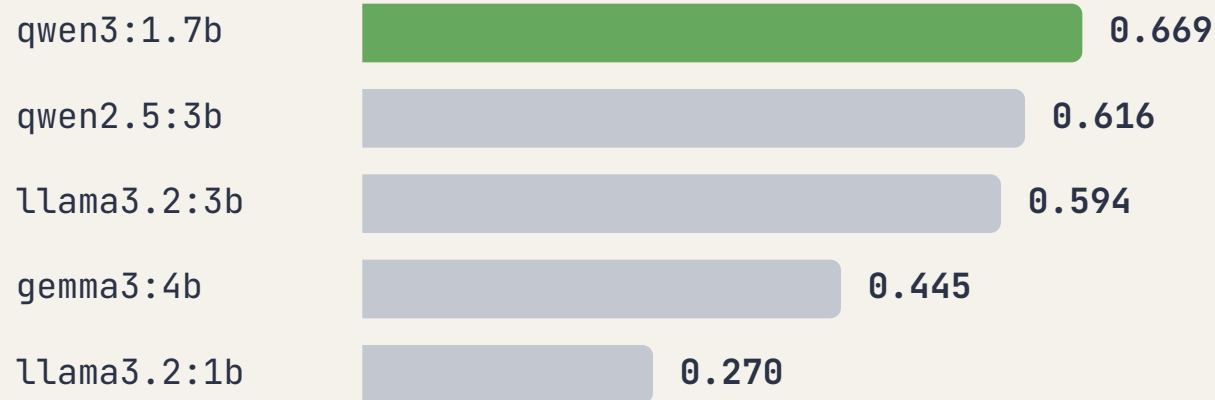
API models

Annotate the full hierarchy, down to individual citations.

SECTION-LEVEL RESULTS

qwen3:1.7b tracks human section rankings best

Spearman ρ with human section scores, mean over 50 papers



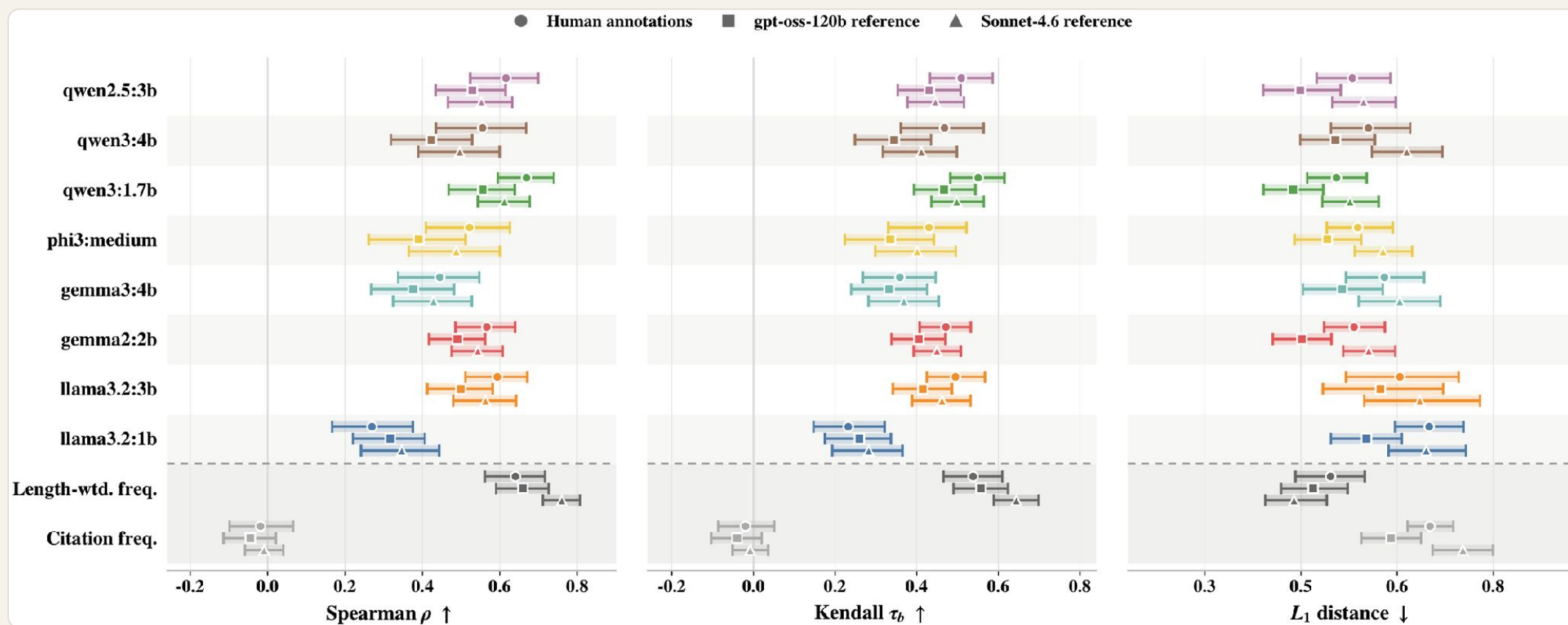
Best LLM on all three metrics against every reference source; lowest L1 vs. humans (0.505).

Length-weighted frequency is strong: best against Sonnet 4.6 ($\rho = 0.760$), but not against humans.

Section length is a strong prior for coarse allocation. Human judgments need content.

SECTION-LEVEL RESULTS

Agreement against all three references



CITATION-LEVEL RESULTS

Frequency heuristics miss the key references

Against Claude Sonnet 4.6, qwen3:1.7b is best on all four citation metrics.

Both frequency baselines show much higher KL and JSD divergence, with heavier tails across papers.

Crediting individual references needs the role each citation plays, not only where and how often it appears.

Paper-level wins for qwen3:1.7b against the 7 other LLMs (max 350, ties excluded)

255

KL divergence

263

JSD

267

Spearman ρ

135

Top-4 overlap

CITATION-LEVEL RESULTS

qwen3:1.7b leads on all four citation metrics

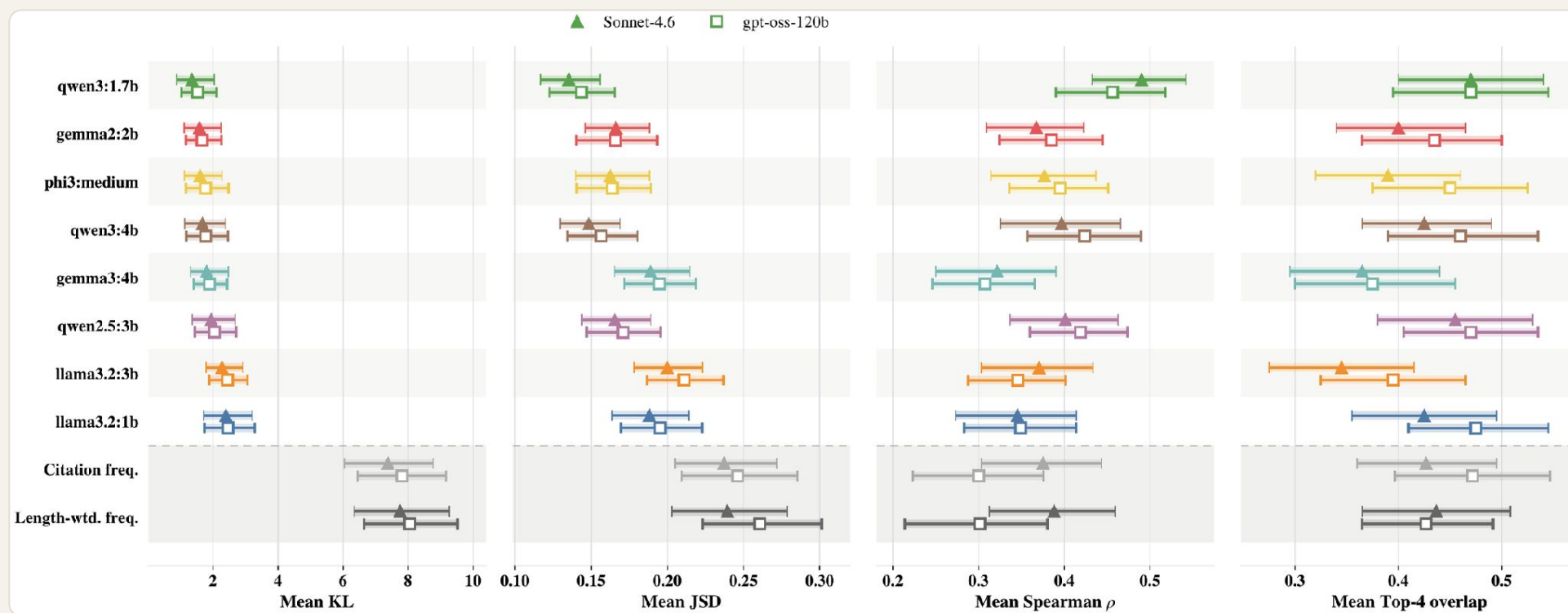


Figure 5. Mean citation-level agreement with Claude Sonnet 4.6 (▲) and GPT-OSS-120B (□) over 50 papers, with 95% bootstrap intervals. Lower KL and JSD, and higher Spearman ρ and Top-4 overlap, are better.

CITATION-LEVEL RESULTS

Per paper, frequency baselines diverge more

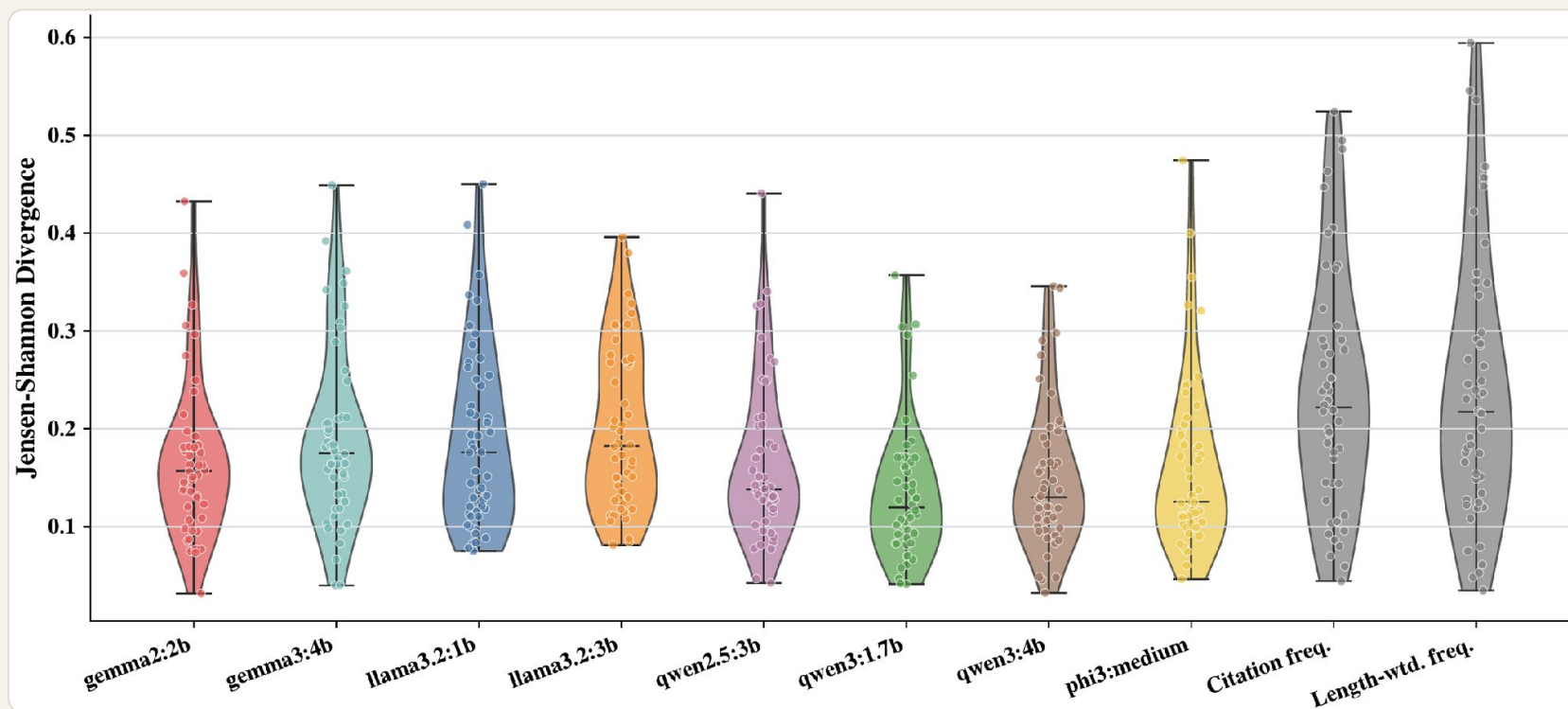


Figure 4. Per-paper Jensen–Shannon divergence from the Claude Sonnet 4.6 citation scores across 50 papers; lower is better. qwen3:1.7b has the lowest divergence, while the frequency baselines are higher and more variable.

ABLATION: MODEL SIZE

Bigger models are not better estimators

Smaller vs. larger	KL	JSD	Spearman ρ	Top-4
qwen3:1.7b vs. qwen3:4b	27–23	27–23	35–15	19–10
llama3.2:1b vs. llama3.2:3b	27–23	30–20	25–25	16–6
gemma2:2b vs. gemma3:4b	29–21	33–17	34–16	16–12

Papers won, smaller – larger model, out of 50 (ties excluded). Citation level, against Claude Sonnet 4.6.

Granularity matters

llama3.2:3b wins at section level, yet 1b wins at citation level.

Caveat

The Gemma pair also differs by generation, so it is not a controlled scaling test.

CORPUS CASE STUDY

Influence, in-degree and contribution diverge

Paper	In-degree	Influence	Contribution
Bahdanau et al. 2014 · Neural machine translation	3	0.475	0.834
Vaswani et al. 2017 · Attention is all you need	6	0.122	0.654
Cheng et al. 2016 · LSTM-networks for reading	1	0.088	0.761
Jozefowicz et al. 2016 · Limits of language modeling	1	0.007	0.755
Kimi Team 2025 · Kimi K2	0	0	0.854

Induced graph of 17 mutually connected papers, scored with qwen3:1.7b. Selected rows.

Bahdanau et al. has 3 in-corpus citations and the largest influence.

Vaswani et al. is most cited, with about a quarter of that influence.

Kimi K2 has the highest contribution but zero influence: nothing here cites it.

TAKEAWAYS

Credit, decomposed

01

The contribution tree conserves credit and separates a paper's original contribution from what it owes its references.

02

Small local LLMs estimate contribution better than frequency heuristics, and scale is not what matters most.

03

Weighted propagation tells apart contribution, cross-paper influence and raw citation count.

Next: instruction-tune compact contribution estimators, and apply the framework to large research corpora.

Thank you!

Check out our codebase for the implementation:

github.com/sanaebrahimi/Importance_Scoring

Check out our paper:

<https://arxiv.org/pdf/2609.07673>

