Auditing Evidence Framing at NeurIPS

Jason Godfrey · University of Michigan

What does progress look like?

A proof.

A formal guarantee.

A better result.

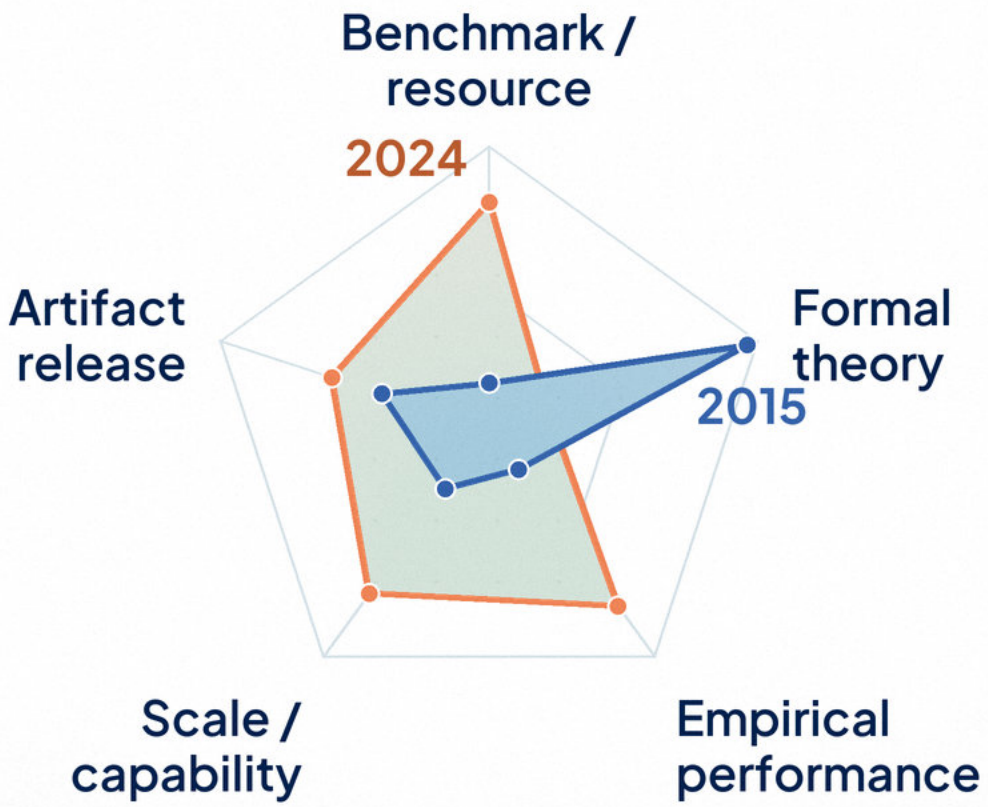
An empirical comparison.

A benchmark.

A resource that others can use.

Abstract

2015 vs. 2024


Benchmark / resource
2024
Artifact release
Formal theory
2015
Scale / capability
Empirical performance

403

accepted papers, NeurIPS 2015

4,497

accepted papers, NeurIPS 2024

19,361

accepted titles and abstracts, 2015 to 2024

Embedding abstract 19,361 of 19,361



100% done (by a computer)

Which evidence does a paper lead with?

theorem 0 matches

“We prove that the estimator converges at the optimal rate.”

- Benchmark / resource
- Formal theory
- Empirical performance
- Scale / capability
- Artifact release

Four positive.
Four negative.

Fixed anchor sentences for each kind of evidence.

Every score compares a paper with fixed anchors

Embed the title and abstract

↓

cos(e, p) – cos(e, n)

↓

Standardize across the corpus → z

Every score compares a paper with fixed anchors

Embed the title and abstract

↓

cos(e, p) − cos(e, n)

↓

Standardize across the corpus → z

Every score compares a paper with fixed anchors

Embed the title and abstract

↓

cos(e, p) − cos(e, n)

↓

Standardize across the corpus → z

Does it show anything?

Checked from several directions

- Anchors
- Encoders
- Lexical probes
- Topic controls

Agreement is not ground truth.

Blinded LLM ratings, with no human gold standard.

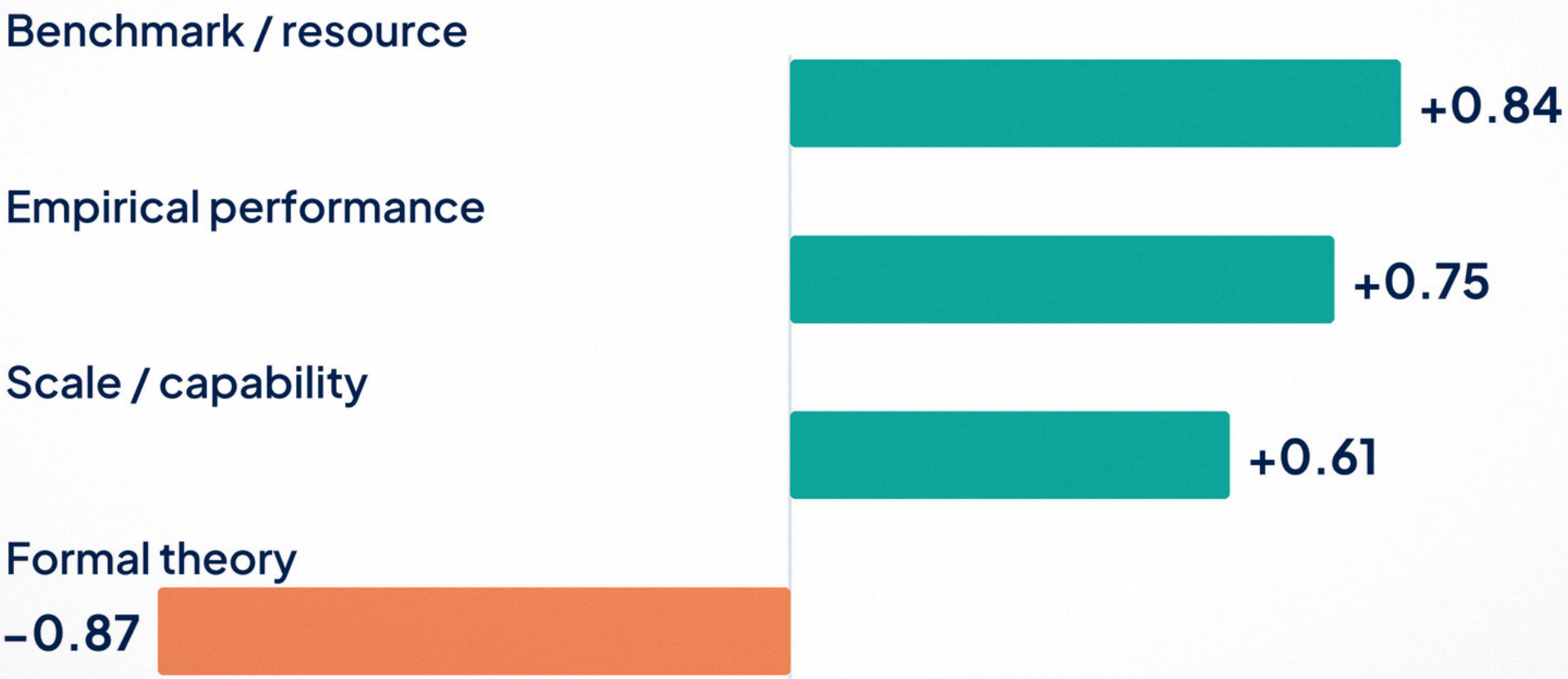
Datasets & Benchmarks track papers

+1.47 SD benchmark framing

−1.22 SD theory framing

What changed from 2015 to 2024?

Three rise. Formal theory falls.



+0.84 SD

benchmark / resource framing, per decade

+0.75 SD

empirical-performance framing, per decade

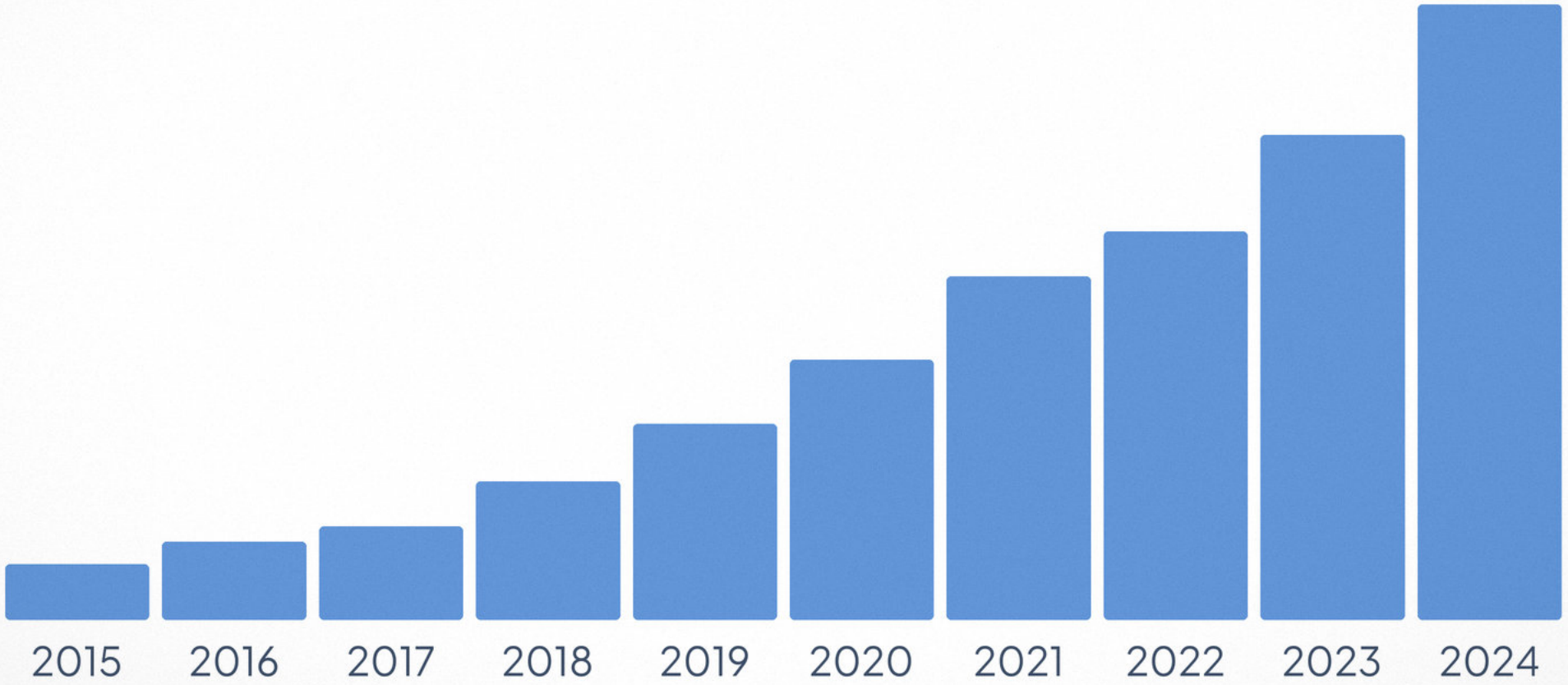
+0.61 SD

scale / capability framing, per decade

−0.87 SD

formal-theory framing, per decade

The conference grew elevenfold.



What they argue about, or how they argue?

40 neighborhoods.

Embedding clusters as topic controls.

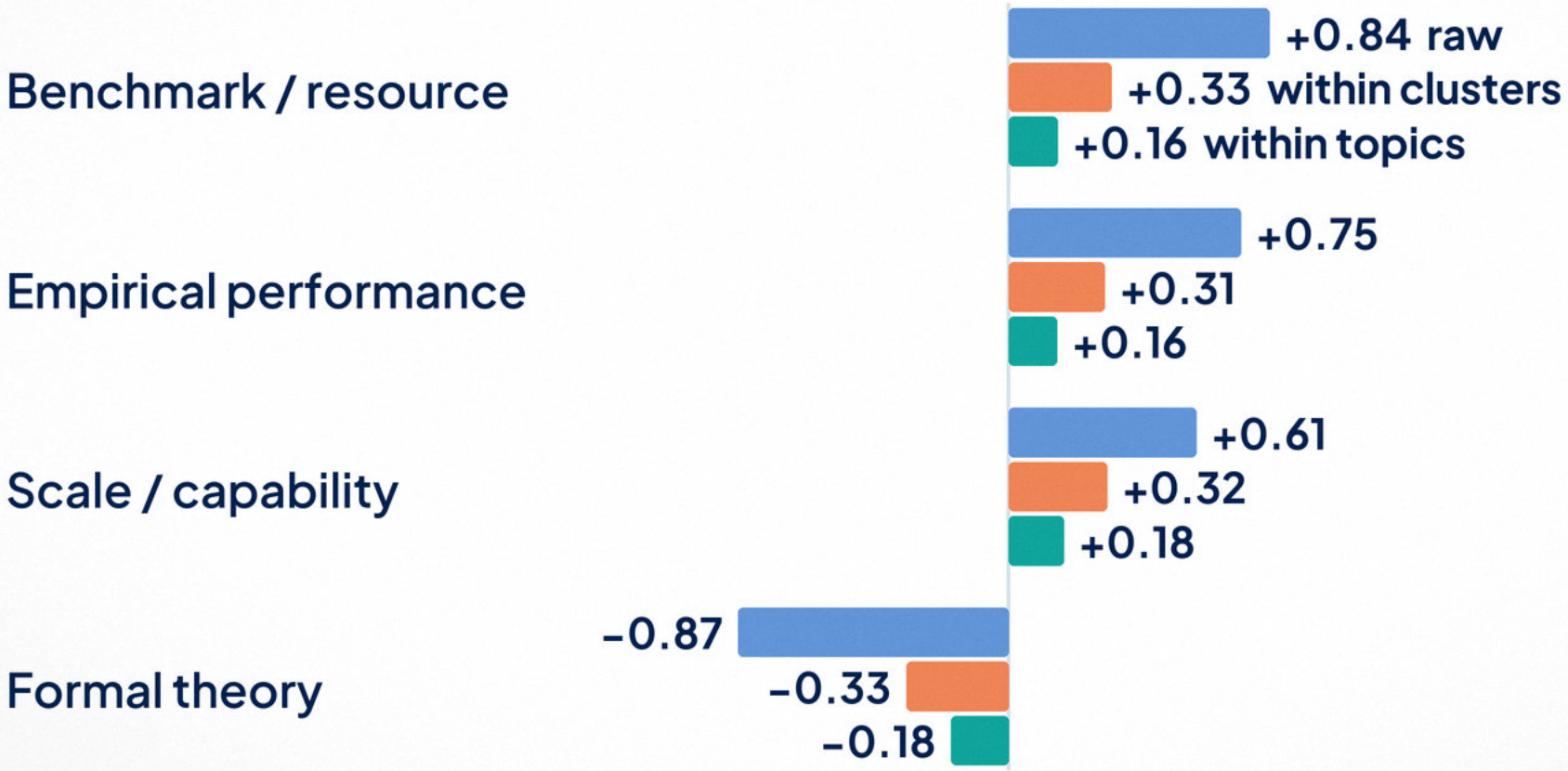
40 lexical topics.

LDA topics from word counts.

Compare papers only with their topic neighbors.

The trends shrink a lot.

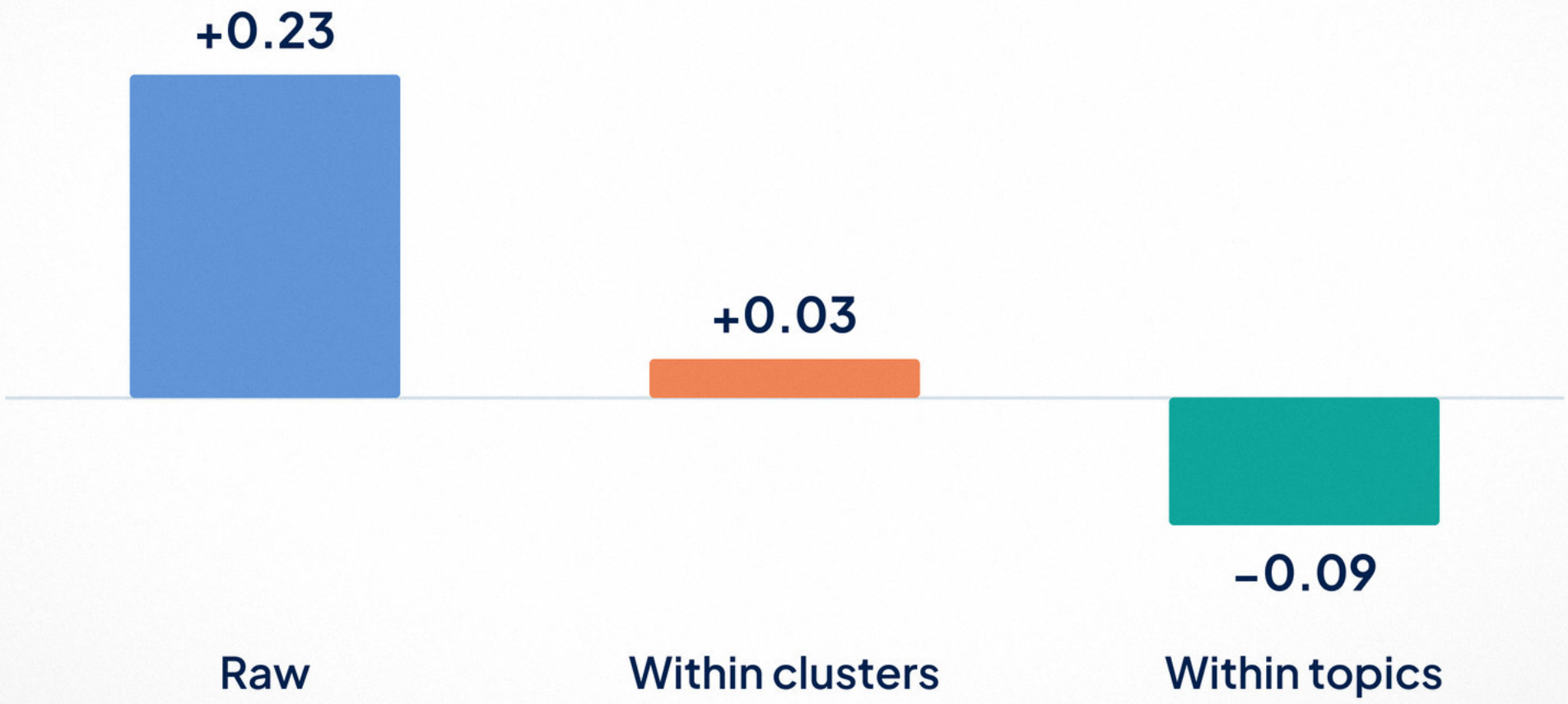
Smaller, same directions.



Same four directions.

Releasing code and data.

The boom that wasn't.



No general open-artifact turn.

What this can’t tell you.

Framing, not quality.

Public summaries.

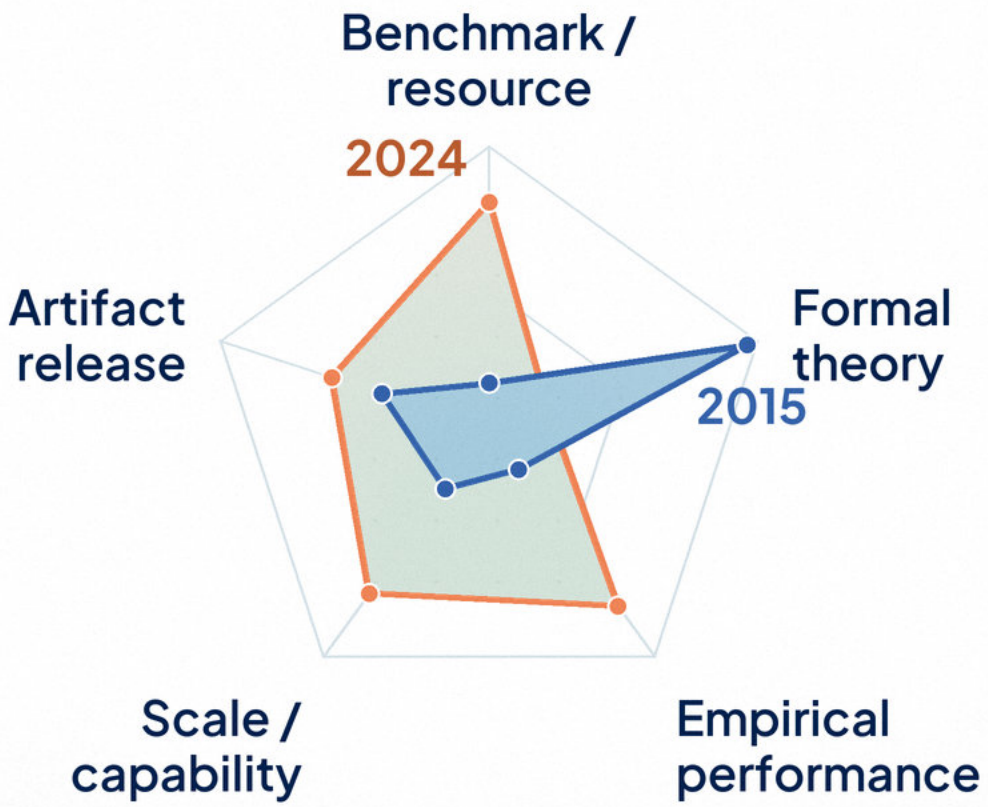
Full papers and rejected
submissions are outside the study.

Reviewer mind-reading

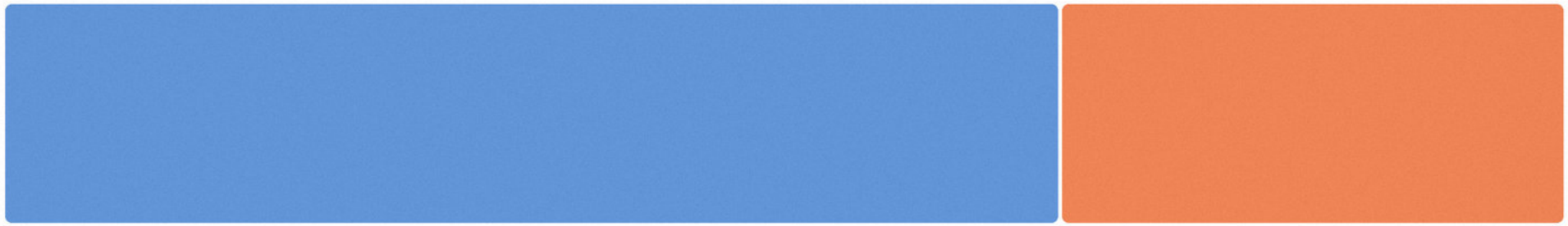
Unavailable in this dataset.

What this can tell you.

2015 vs. 2024


Benchmark / resource
2024
Artifact release
Formal theory
2015
Scale / capability
Empirical performance

Topics explain much, not all.



+0.52 from the topic mix

+0.25 within topics

How do NeurIPS papers argue for progress?

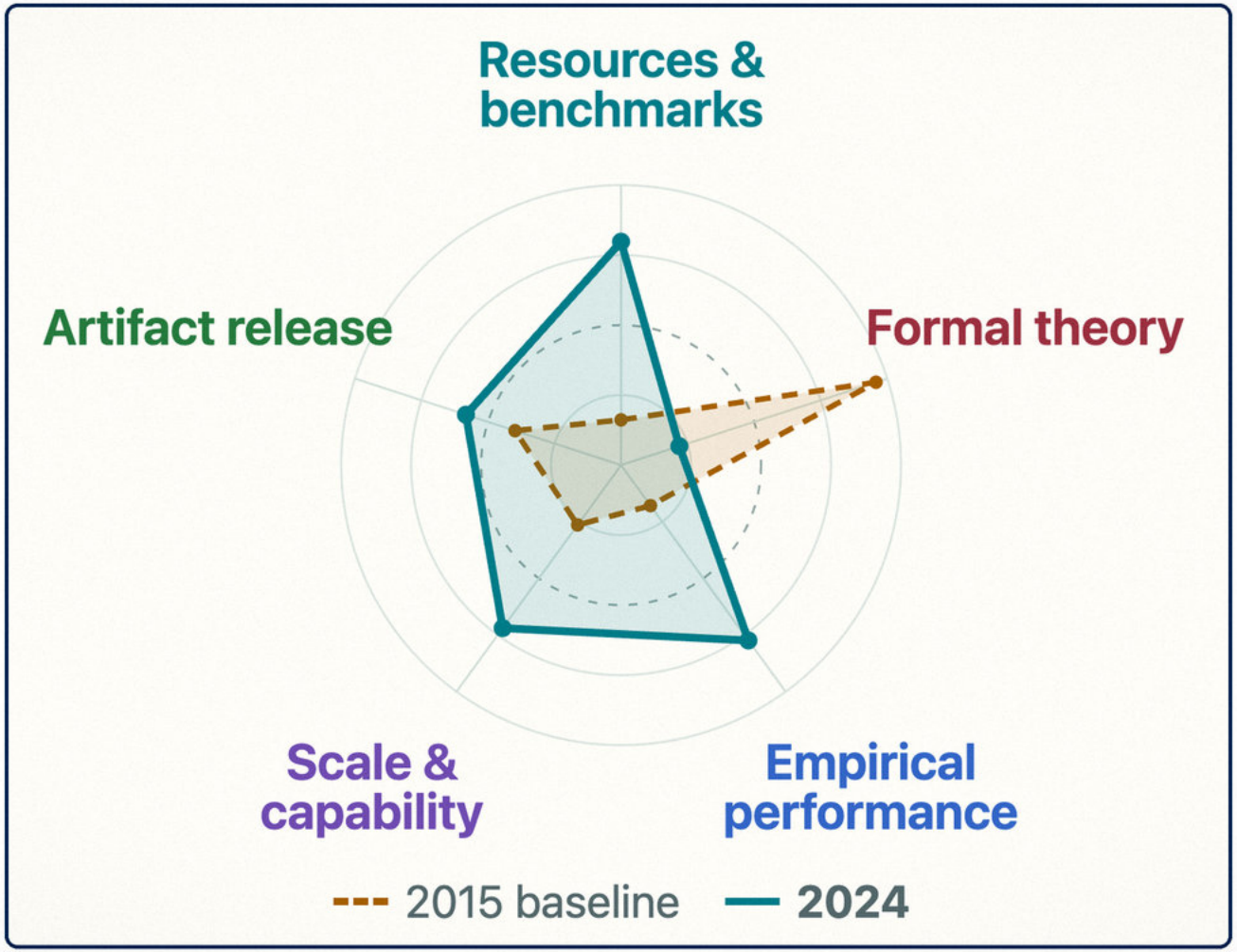
Rarely measured. Now there is a baseline.

Extend.

Other venues and later years need new validation.

Reproduce.

Data, frozen anchors and code are on Kaggle.


Resources & benchmarks
Formal theory
Empirical performance
Scale & capability
Artifact release
2015 baseline
2024

Not a leaderboard.

benchmark

1,059 matching papers · all years

∇²DFT: A Universal Quantum Chemistry Dataset of Drug-Like Molecules and a Benchmark for Neural Network Potentials

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin,...

2024 Proceedings ↗

3DCoMPaT200: Language Grounded Large-Scale 3D Vision Dataset for Compositional Recognition

Mahmoud Ahmed, Xiang Li, Arpit Prajapati, Mohamed Elhoseiny

2024 Proceedings ↗

Compare.

Try other anchors and see what changes.

Machine learning changed fast.

Homework

J. Godfrey



Explore the evidence.

jasongodfrey.com/atlas

