More Computational Resources Do Not Ensure Higher Scholarly Impact:

Evidence from Leading NLP Conference Papers

Shuai Chen, Tong Bao, Jitong Peng, Chengzhi Zhang*

Nanjing University of Science and Technology

Why Study Compute and Scholarly Impact?

- **Compute expands experimental possibilities.**
  Larger models and demanding experiments require substantial resources.

- **Access to compute is uneven.**
  Researchers and institutions face different resource constraints.

- **Experimental capability and scholarly impact are different outcomes.**
  We examine citations and paper awards—not model performance.

Are greater reported GPU resources associated with higher scholarly impact?

Related Work: Beyond Model Performance

1. Scaling and model performance → 2. Costs, reporting, and access → 3. Compute and scholarly impact

- Scaling laws relate **model performance** to compute, model size, and data.
- Compute-reporting research emphasizes **efficiency, transparency, and reproducibility**.
- Compute-divide studies document **unequal access** across institutions and sectors.
- Foundation-model studies link GPU resources to **publication and citation outcomes**.

Kaplan et al., 2020; Strubell et al., 2019; Schwartz et al., 2020; Ahmed & Wahed, 2020; Besiroglu et al., 2024; Hao et al., 2025

Three Research Questions

1. Resource concentration versus impact concentration

Do the highest-capability papers account for a comparable share of citations and awards?

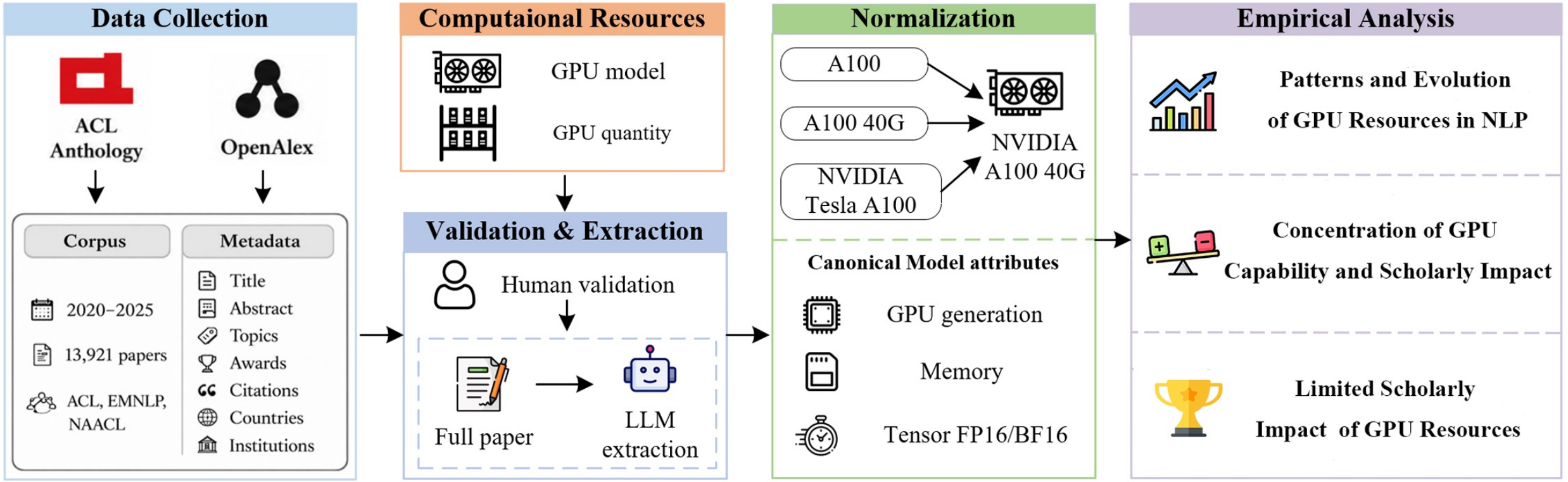
2. Associations after adjustment

Do associations persist after accounting for publication context, topic, and team structure?

3. GPU count versus hardware generation

Do deployment scale and newer hardware relate differently to scholarly impact?

Study Design



Corpus	Reported GPU capability	Measurement boundary
13,921 main-conference papers ACL · EMNLP · NAACL 2020–2025	Largest reported configuration GPU count × per-GPU peak throughput	Hardware capability, not consumption No runtime or utilization adjustment

Data Collection & Enrichment

1. **Corpus:** ACL, EMNLP, and NAACL main-conference papers, 2020–2025

2. **Full text:** parse accessible PDFs with MinerU

3. **Metadata:** link citations, team information, and normalized affiliations

4. **Research context:** official awards + one of 29 primary NLP topics

5. **Final corpus:** 13,921 papers

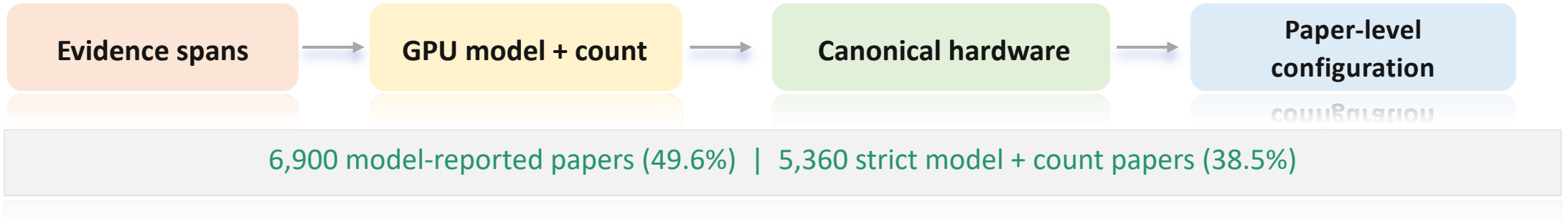
GPU Resources: Extraction & Validation

1. **Test set:** 400 manually annotated papers;
   120 independently double-annotated after guideline refinement

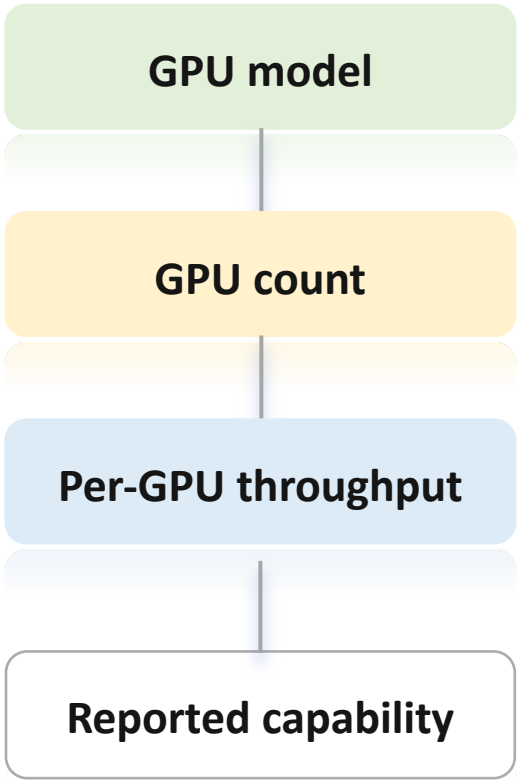
2. **Evidence identification:** $\kappa = 0.94$
   Exact agreement: GPU model 90.83%; GPU count 87.50%

3. **DeepSeek-v3.2:** model F1 = 0.933; model + count F1 = 0.879
   Selected by balancing extraction accuracy and processing cost

4. **Hardware normalization:** standardize model names, variants,
   generations, and peak Tensor FP16/BF16 throughput



GPU Resources: Measuring Capability



Largest reported GPU configuration

$$GPU_Capacity_i = \max \text{ over reported configurations } (\text{GPU count} \times \text{per-GPU peak Tensor FP16/BF16 throughput})$$

- Retain the **largest configuration**, rather than summing across experiments.
- Measure **reported capability**, not GPU-hours, training FLOPs, cost, energy, or utilization.

Example: 8 × A100 → **8 × per-GPU peak rate**

Model-only reports: count = 1 is a lower bound. **Non-reporting ≠ zero capability.**

Impact: Measures & Adjusted Models

Citation / award outcomes

Normalize citations

Adjust for research context

Conditional associations

Citation outcomes

Primary	Within-NLP topic–year citation percentile
Secondary	OpenAlex field-normalized percentile
Additional	log(1 + citations), citation counts, top-10% status

Award outcome: official paper-award indicator

Baseline controls

Year × venue and primary-topic fixed effects

Team size and organization count

Robustness: expanded author / institutional controls

Citations: 2020–2023 strict sample, N = 2,194 | Awards: 2020–2025, N = 5,357

Analysis Roadmap

Describe reported resources, compare impact, then adjust for research context.

1. Reporting and scale

- ✓ Reporting completeness
- ✓ GPU count and generation
- ✓ Reported capability over time

2. Research context

- ✓ Institutional settings
- ✓ Primary research topic
- ✓ High-capability group

3. Resource–impact concentration

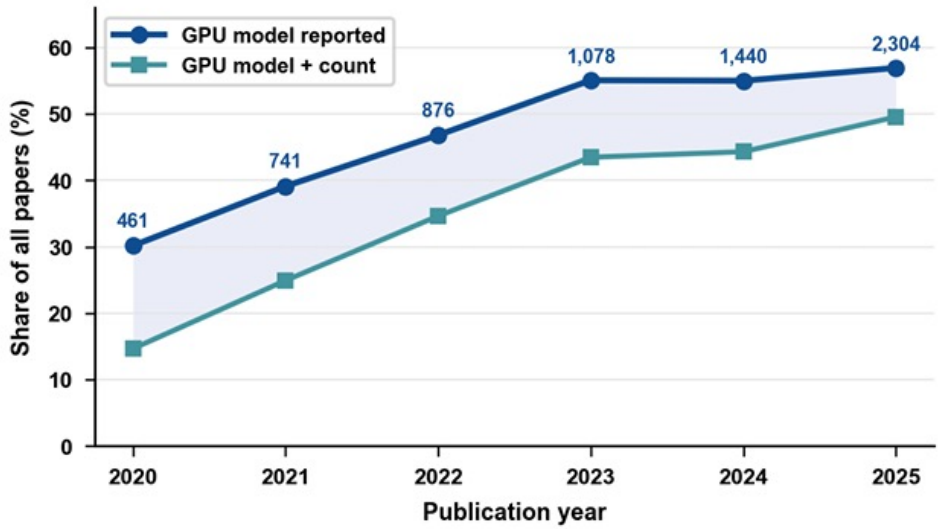
- ✓ Annual top-20% capability share
- ✓ Citation and award shares
- ✓ Overlap with highly cited papers

4. Adjusted associations

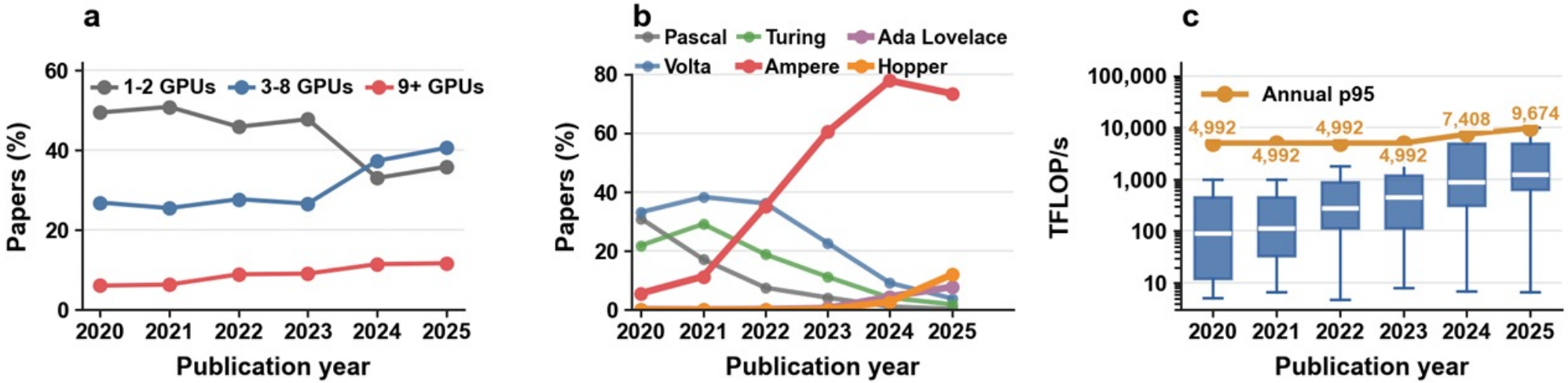
- ✓ Continuous capability models
- ✓ GPU count versus generation
- ✓ Expanded controls and Findings

Results: Reporting Improved and Capability Rose

- GPU model disclosure: **~30% → 57%**
  Model + count disclosure: **~15% → 49%**
- 1–2 GPU configurations: **49.5% → 35.9%**
- Median reported capability: **~91 → 1,248 TFLOP/s**
- A100-class GPUs most common from **2023**

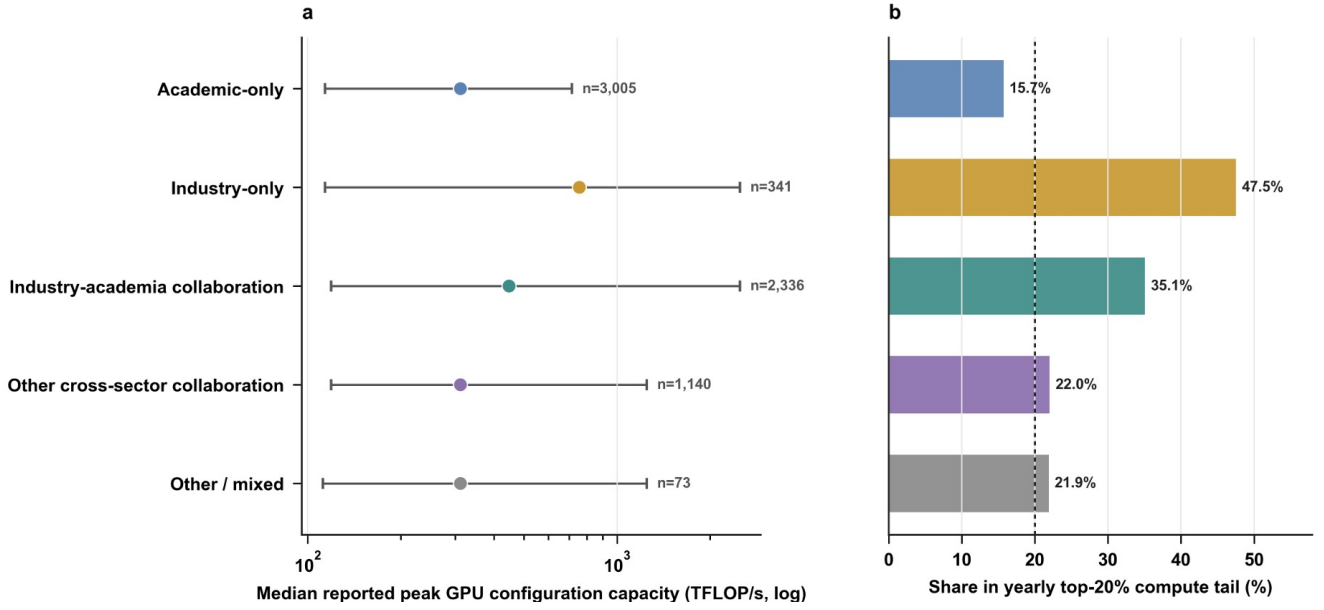


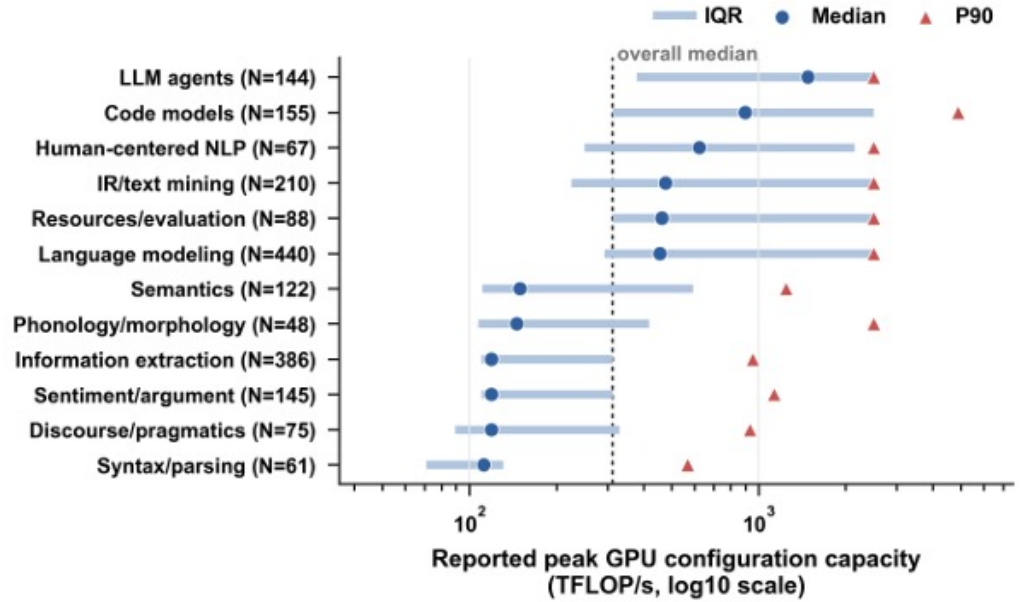
Growth mainly reflects newer hardware and medium-scale configurations, not very large clusters.



Results: Capability Varies by Institution and Topic

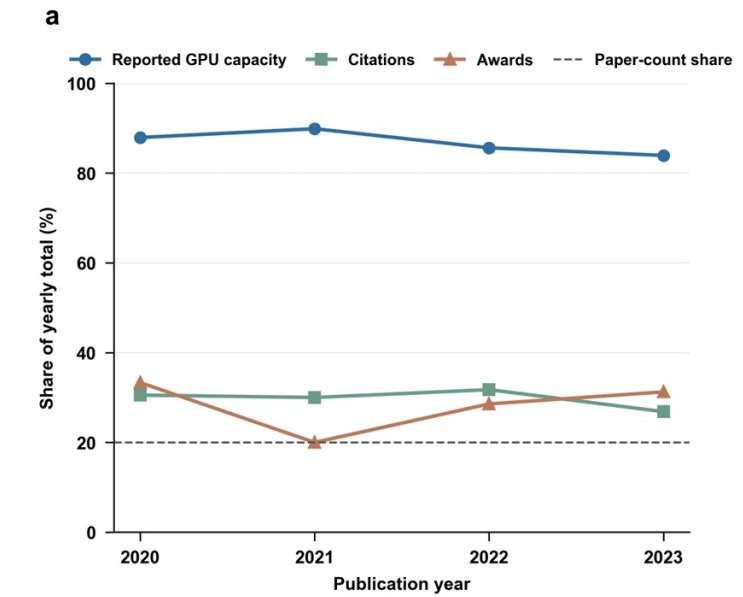
- Industry-involved papers report **higher median capability** and appear more often in the annual top-20% capability group.
- Higher-capability configurations cluster in **LLM agents, code models, and language modeling**.
- Research context therefore matters when interpreting compute–impact associations.

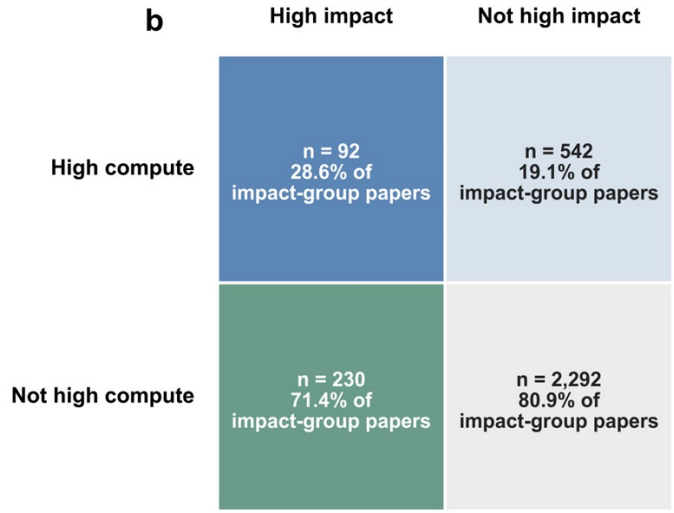




Results: Resource Concentration Exceeds Impact Concentration

- The annual top 20% account for **83.9–89.9% of reported GPU capability**.
- The same papers account for **27–32% of citations** and **20–33% of awards**.
- Citation top-10% rate: **14.5% vs. 9.1%** for high-capability versus other papers.
- Yet **85.5%** of high-capability papers are outside the citation top 10%.





2020–2023 model-reported sample (N = 3,156). High capability: annual top 20%; high impact: citation top 10% within year × venue. Shares use the GPU-quantifiable sample.

Results: Adjusted Associations Are Positive but Limited

- 10× reported capability is associated with **+3.52 pp** in the primary citation percentile (95% CI: 1.27–5.77); incremental **$R^2 = 0.0042$**.
- **GPU count** shows the more consistent citation and award associations; newer hardware has no comparably robust award association.
- **Expanded controls and a Findings replication** retain the broad pattern of limited incremental explanatory power.

Outcome	GPU capability (per 10×)	GPU count (per 10×)	Ampere or newer (vs. earlier)	ΔR^2 aggregate / joint
NLP topic–year percentile (primary)	**+3.52** pp	**+4.57** pp	**+4.16** pp	**0.0042 / 0.0093**
OpenAlex percentile (secondary)	+1.26 pp	+1.96 pp	+2.02 pp	0.0009 / 0.0033
log(1 + citations), OLS	+18.8%	+24.7%	+15.0%	0.0052 / 0.0086
Citation count, PPML	+61.2%	+69.3%	+31.0%	—
Top-10% cited, LPM	+3.84 pp	+4.41 pp	+2.17 pp	0.0043 / 0.0051
Awarded, LPM	+0.86 pp	+1.33 pp	−0.25 pp	0.0011 / 0.0018

Count and generation are estimated jointly in a separate model, excluding aggregate capability.
pp = percentage points. * p < .05; ** p < .01; *** p < .001. Citation N = 2,194; award N = 5,357.

Limitations & Future Work

- **Measurement:** reported configurations, not realized consumption.
  Runtime, utilization, and cumulative experimental use remain unmeasured.

- **Coverage:** non-reporting and API-mediated compute limit visibility.
  No GPU report does not mean compute-free research.

- **Interpretation:** observational associations, not causal effects.
  Citations and awards are incomplete proxies for scholarly value.

- **Next steps:** standardized consumption reporting, external compute,
  and broader coverage of journals, workshops, and other research settings.

Support access to research infrastructure—
without treating hardware scale as a substitute for scientific contribution.

Thank you!



Paper

EMNLP 2026 • Main Conference