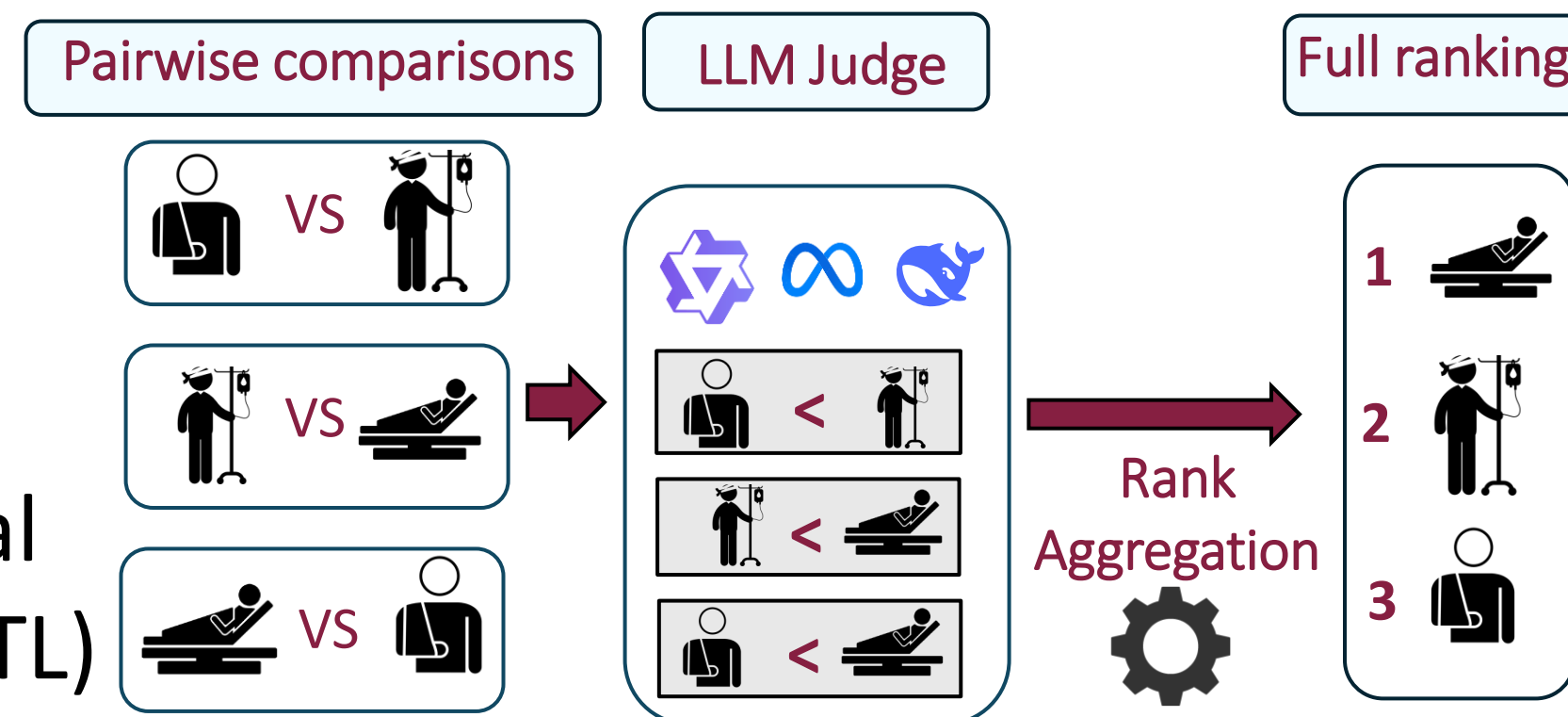


Motivation

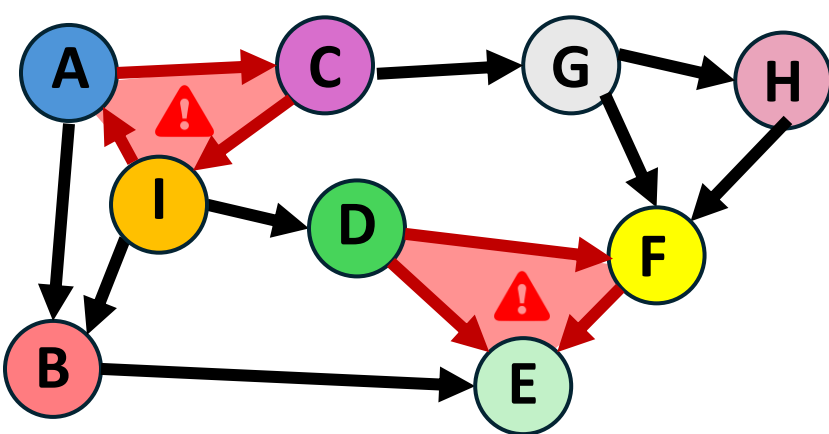
- LLMs are increasingly being used for ranking tasks
- However, directly asking the LLM to rank a list of items is difficult:
 - Sensitivity to presentation order, list length, prompt framing, etc.
 - Error-prone and unstable

- Alternate method backed by social choice theory:
 - Elicit pairwise comparisons
 - Aggregate pairwise choices to a total order (using Rank Centrality[4] or BTL)

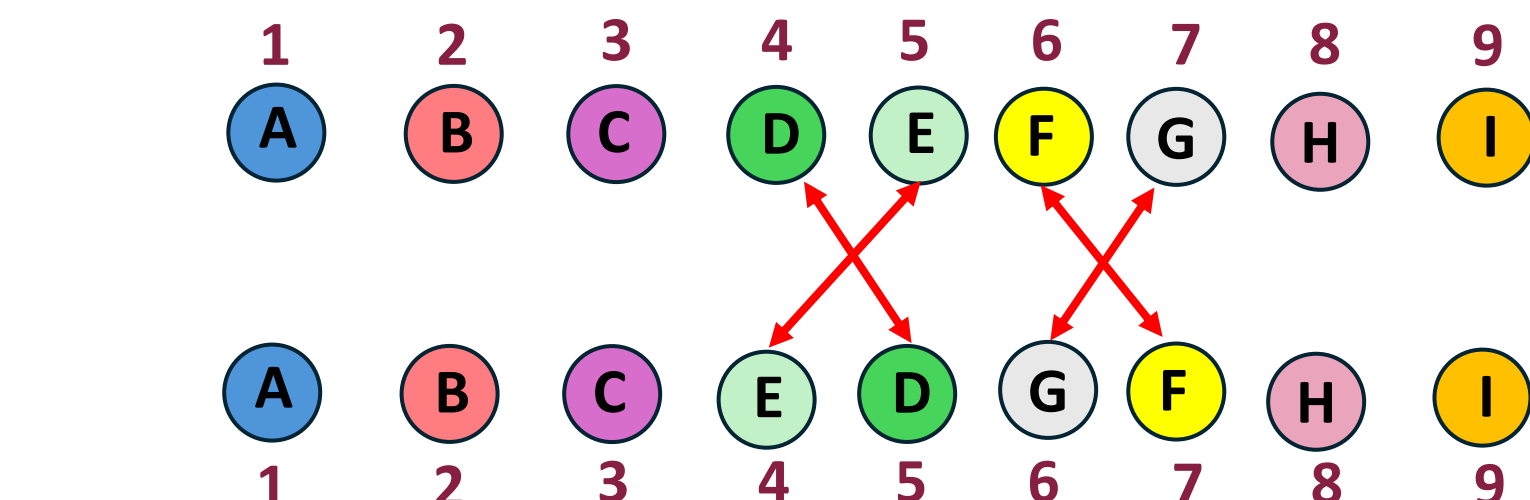


But how to assess reliability of such rankings?

Two approaches in the literature



1. Intra-ranking consistency: # of transitivity violations when looking at the full set of pairwise comparisons (e.g. coefficient of consistency (ζ) [1])



2. Inter-ranking consistency: Stability or variance of the ranking outcome. Compare two different rankings generated by two different runs of the LLM, and compare consistency in the final ranking (for example: Kendall's τ [5])

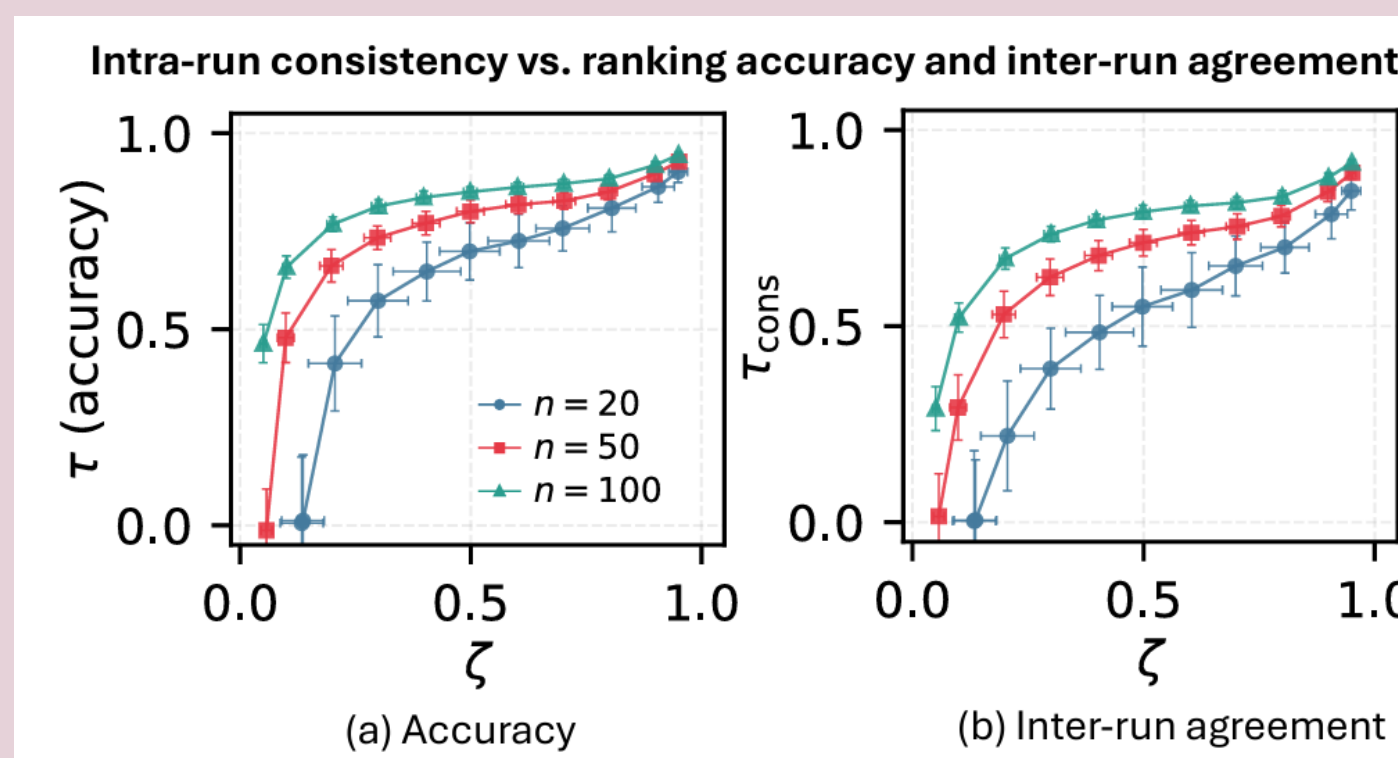
What is the empirical relationship between these 2 metrics?

- Especially when such LLM generated rankings inform consequential decision-making on real world problems

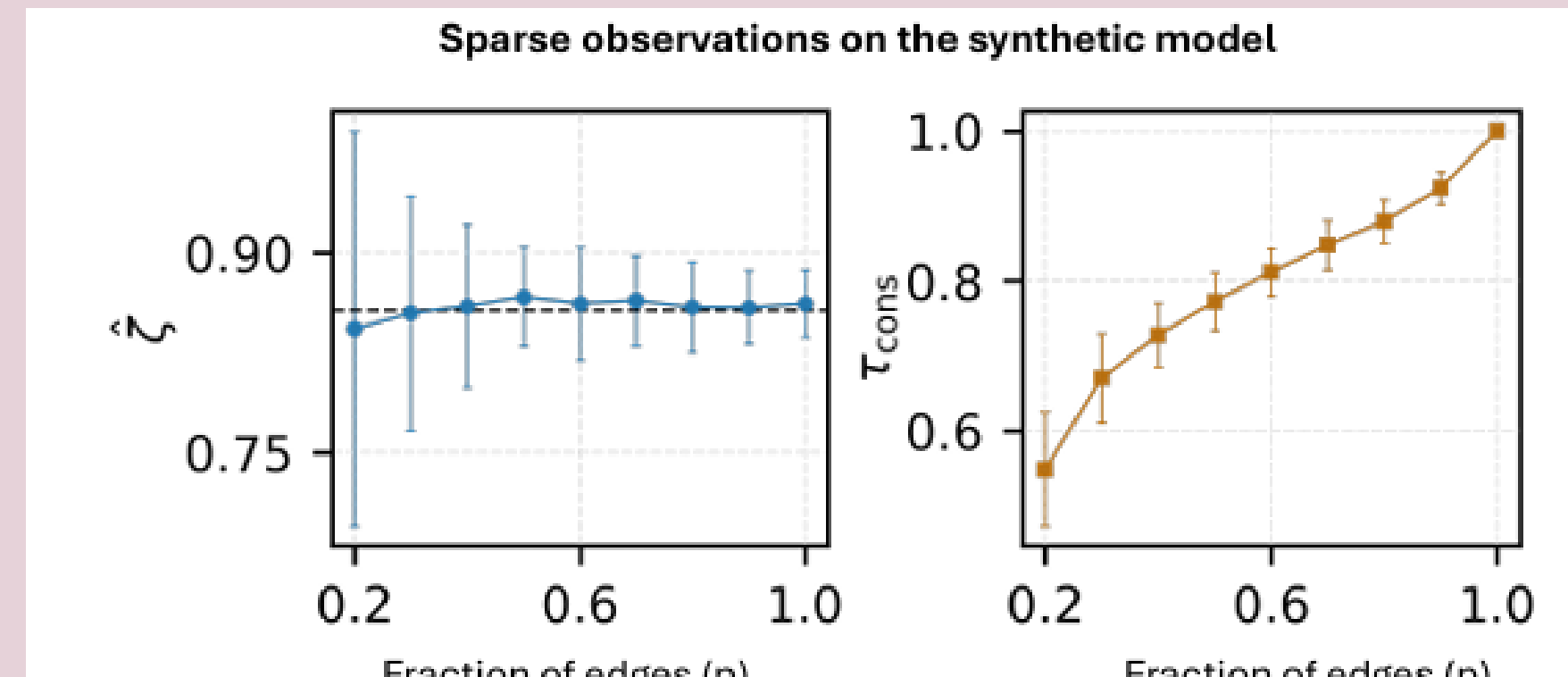
		Inter-run variance (IV)	
		Low	High
Intra-run consistency (C)	High	Borda, consistent triads	Random-permutation
	Low	Fixed-cycle	Noisy method, inconsistent triads

ζ and τ in Synthetic Bradley-Terry-Luce generated rankings

- From a synthetic known ranking, we generate noisy pairwise comparisons using the BTL model:
 - For 2 items i and j , their pairwise comparison is determined by $P(i > j) = \frac{1}{1 + \exp(-\beta(j-i))}$; where $\beta \geq 0$ governs noise.
- From comparisons, the full ranking is reconstructed using Rank Centrality (RC) [4]



Intra-run consistency compared with inter ranking consistencies. τ (accuracy) is Kendall's τ between the reconstructed ranking and known true ranking. τ_{cons} is Kendall's τ between two independent RC ranking from the same generative model. We vary the size of the ranking (n). At a given ζ , larger tournament size (n) yields better inter-run agreement. **Inter rank consistency (τ) depends on both intra-run consistency (ζ) and tournament size.**



Sparse Comparisons: For $n = 30$ items, p fraction of possible comparisons are performed.
Takeaways:

- Even with sparse comparisons, ζ remains relatively constant and a good estimator of intra-run consistency (ζ)
- Inter-run consistency (τ_{cons}) increases monotonically with p .

$\hat{\zeta}$ is a good measure of comparison consistency.
 τ_{cons} is a signal of both comparison consistency and quantity.

Evaluating LLMs on real data from Homelessness Resource allocation and Emergency Room Triage

Dataset	Available records	Items used
VI-SPDAT [2]	325 households	325 households
VI-F-SPDAT [6]	698 households	698 households
TAY-VI-SPDAT [7]	561 households	561 households
MIMIC-IV [3]	180,057 patients	500 patients

Theoretical analysis carries over to LLMs' performing prioritization tasks in real-world tasks

- Prioritizing households for receiving homelessness resources: VI-SPDAT, VI-F-SPDAT, TAY-VI-SPDAT
- Prioritizing patients for medical emergency department triage: MIMIC

We prompt:

- 3 LLMs (Qwen2.5-7B-Instruct, Meta-Llama-3-8B-Instruct, and DeepSeek-R1-Distill-Llama-8B)
- To do pairwise prioritization (choose household A vs B)
- Use RC to aggregate comparisons into a ranking

We sampled edges in the comparison graph with $p=0.4$. Each "comparison sample->prompt->rank aggregation" pipeline was performed twice, yielding 2 different rankings for each (LLM, dataset) pair.

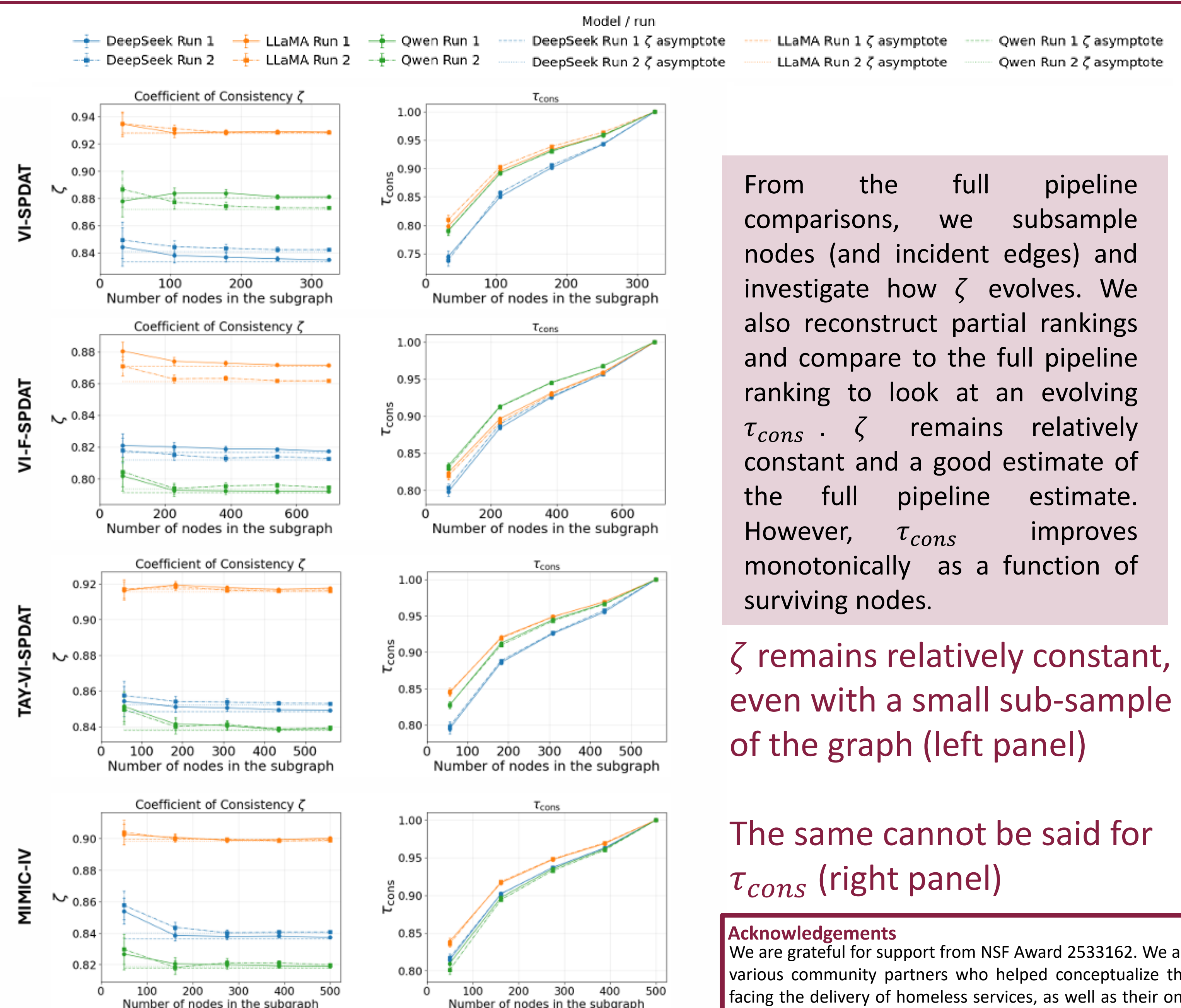
ζ and τ are measurements of consistency on orthogonal axes

Measure	Consistency axis	Comparison budget	Aggregation method
ζ	Pairwise comparison consistency	Independent in expectation	Independent
τ	Ranking consistency	Dependent	Dependent

Both are informative and necessary!

Dataset	DeepSeek		LLaMA		Qwen	
	ζ	τ_{cons}	ζ	τ_{cons}	ζ	τ_{cons}
TAY-VI-SPDAT	○	—	+	○	—	+
VI-F-SPDAT	○	○	+	—	—	+
VI-SPDAT	—	—	+	○	○	+
MIMIC-IV	○	○	+	—	—	+

Relative model ranking by average ζ and Kendall rank correlation (τ_{cons}). + indicates best, ○ middle, and — worst among the three models for that dataset and measure



From the full pipeline comparisons, we subsample nodes (and incident edges) and investigate how ζ evolves. We also reconstruct partial rankings and compare to the full pipeline ranking to look at an evolving τ_{cons} . ζ remains relatively constant and a good estimate of the full pipeline estimate. However, τ_{cons} improves monotonically as a function of surviving nodes.

ζ remains relatively constant, even with a small sub-sample of the graph (left panel)

The same cannot be said for τ_{cons} (right panel)

Acknowledgements
We are grateful for support from NSF Award 2533162. We also thank the various community partners who helped conceptualize the challenges facing the delivery of homeless services, as well as their ongoing efforts to support local families.

References
[1] Kendall, M. G., & Babington Smith, B. (1940). On the method of paired comparisons. *Biometrika*, 31(3–4), 324–345.
[2] OrgCode Consulting Inc., & Community Solutions. (2015). *Vulnerability Index–Service Prioritization Decision Assistance Tool (VI-SPDAT): Prescreen triage tool for single adults*.
[3] Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). *MIMIC-IV* (Version 3.1) [Data set]. PhysioNet.
[4] Negahban, S., Oh, S., & Shah, D. (2017). Rank Centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1), 266–287.
[5] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1–2), 81–93.
[6] OrgCode Consulting Inc., & Community Solutions. (2015). *Family Service Prioritization Decision Assistance Tool (F-SPDAT): U.S. version 2.0*.
[7] OrgCode Consulting Inc., Corporation for Supportive Housing, Community Solutions, & Rice, E. (2015). *Next Step Tool for Homeless Youth (TAY-VI-SPDAT): U.S. version 1.0*.