Interactive Steering of LLM judges with ATTUNE

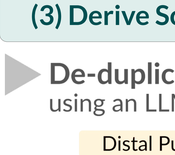
Bhavya Chopra , Meng Chen, Rebecca Dang, Chanbin Park, Shreya Shankar, Sepanta Zeighami, Björn Hartmann, Aditya G. Parameswaran

Contact: bhavyachopra@berkeley.edu

Prompt-based LLM Scoring is Brittle

Lack of Holistic Data Understanding

- Records scored in isolation may lead to inconsistent judgments.



Score each patient on the severity of their injuries on a scale of 1 (least severe) to 4 (most severe)

Patient X
Bruised arms, legs
Score: 3

Patient Y
Airway compromise
Score: 3

Lack of Interpretability

- Prompt to score mappings are opaque and illegible
- Post-hoc explanations may be insufficient and do not scale!

Lack of Steerability

- Natural language prompts are the only control mechanism
- Prompt edits don't ensure improved LLM judgments

ATTUNE: Consistent Scoring Algorithm

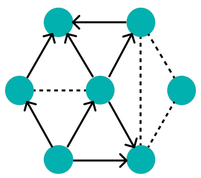
(1) Gather Holistic Data Understanding

Perform Comparisons
between text inputs using LLM Judge

13 y/o M, bleeding finger → Poor Circulation → 7 y/o M, RR 38, absent distal pulse

"Prioritize 7 y/o over 13 y/o due to poor circulation"

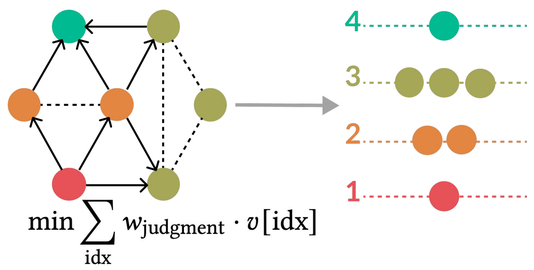
Construct Comparison Graph
where edges denote pairwise preferences or ties



(2) Solve Scoring as an ILP

Assign Consistent Scores
by minimizing inconsistencies in assigned scores and pairwise preferences:

- If 1 → 2, then score(2) > score(1)
- If 1 ⇢ 2, then score(1) = score(2)



(3) Derive Scoring Criteria Bottom-Up

De-duplicate Criteria
using an LLM

Distal Pulse, Circulation, Non-palpable → Capillary Refill

Rank Criteria
based on frequency

Circulation, Unresponsive, Nausea → Unresponsive, Circulation, Nausea

Map Criteria
Evaluate each criterion on all inputs to obtain a criteria matrix

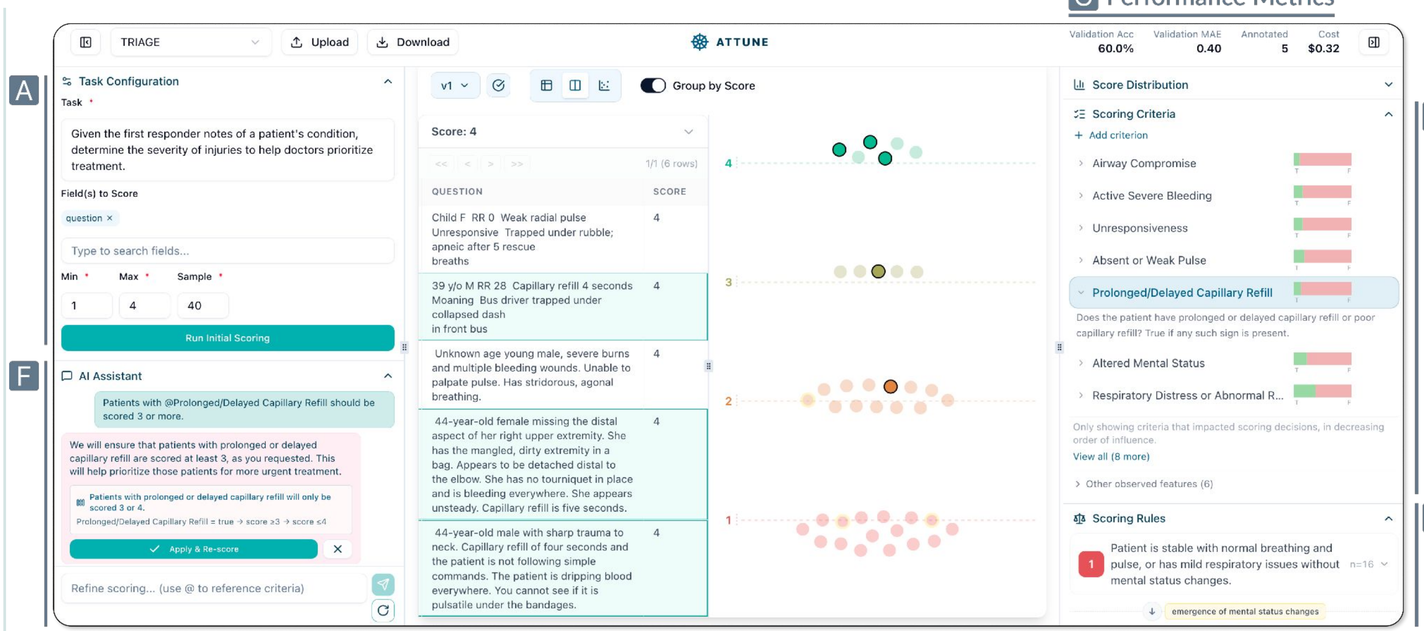
(4) Discover Rules via Greedy Set Cover

Construct Scoring Rules
by identifying similar conditions within each score group

3 ⇢ (Capillary Refill > 2 seconds **AND** Altered Mental State = True) **OR** (Abdominal Injury = True)

Generate NL Summary
of scoring rules and decision boundaries using an LLM

"Patients scored 3 may exhibit altered mental state, signs of shock, or severe abdominal injuries, but are responsive"



ATTUNE Improves Steerability and helps users Co-Evolve Logic

Users quickly trusted scoring criteria derived bottom-up.

"I like that the criteria are curated based on actual essay contents. When I go in without a rubric, or with one that was created upfront and handed down to me, it is easy to be harsh for the first couple essays.

Scoring criteria and rules let users catch misalignments *at-a-glance*

"I should not be seeing patients with capillary refill delays under score 2. Let me fix this.

Users *show and tell*: interactions spanned direct edits, chat, and combinations

"It's easier to select items and tell something in the chat when I want the rules to be different

- **Task-based user study (N = 8)**
- **Domain experts** in healthcare, law, AI evals, and education
- **60 minute sessions**
- **Qualitative analysis** using grounded theory

	Seed		Examples		Distribution		Rules	
	Prompt	ATTUNE	Prompt	ATTUNE	Prompt	ATTUNE	Prompt	ATTUNE
Patient Triage	0.36	**0.60**	0.46	**0.63**	0.44	**0.58**	0.78	**0.86**
Student Essays	0.08	**0.14**	0.32	**0.51**	0.44	**0.52**	0.58	**0.70**
Product Relevance	0.39	0.32	0.33	**0.46**	0.32	**0.61**	0.53	**0.62**

- Technical evaluation with **3 workloads**
- **Baseline:** GEPA optimized prompts
- **ATTUNE beats prompt-based scoring with identical guidance**