

Agreement-Aware Selective Prediction for Museum Label Transcription

Md Raihan, Mathias Zinnen, and Vincent Christlein

Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Selective prediction for museum label transcription

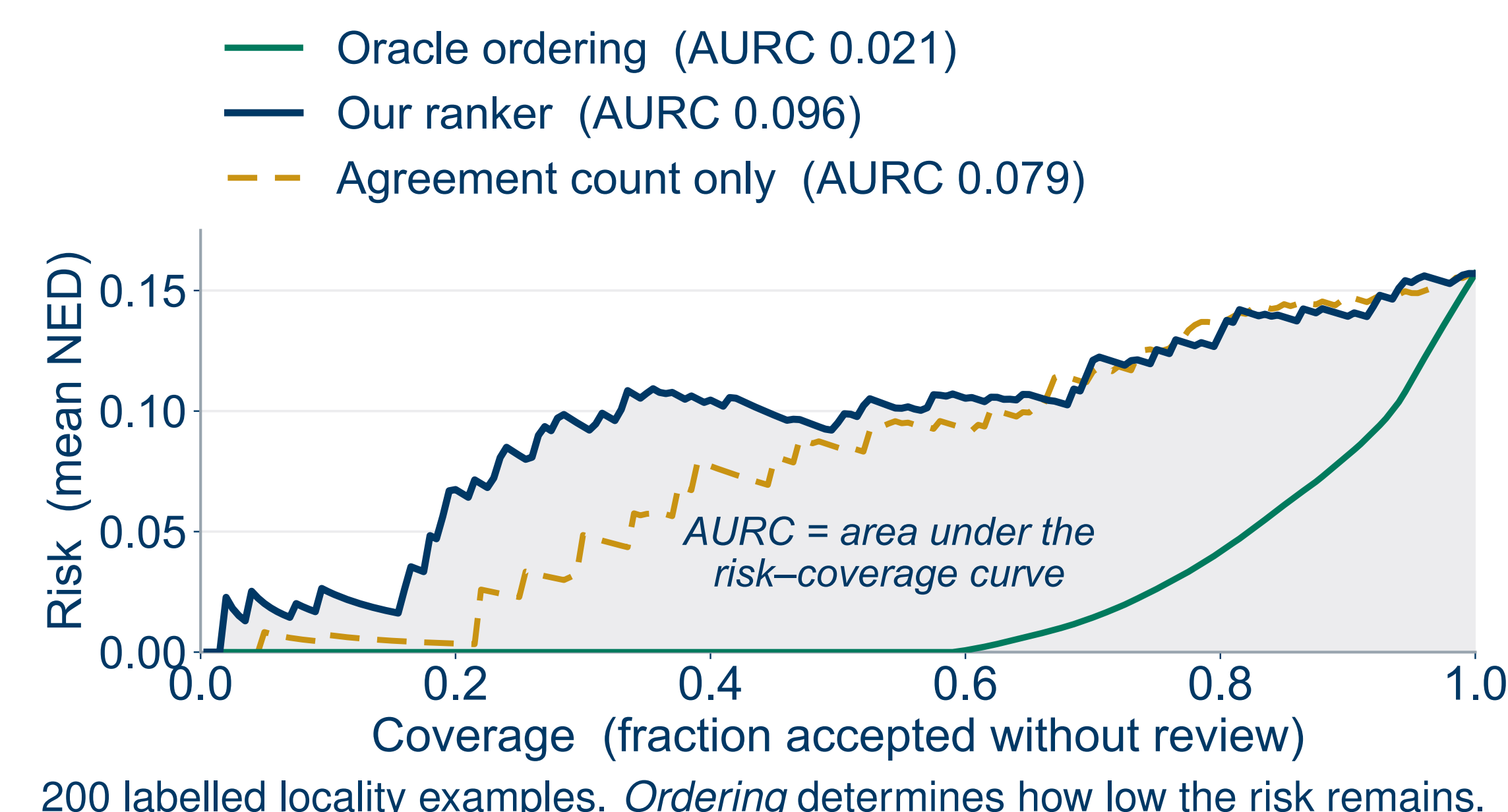
- Natural-history collections are digitised **far faster than their labels can be read** [2]: handwritten, abbreviated, archaic, and often **split across several cards**.
- We study whether prediction disagreement can identify samples that should be sent for curator review.
- MuseumSCAT 2026** [1] provides the Darwin Core [3] fields `verbatimDate` and `verbatimLocality` for **3,300** photographs of pinned Danish dung beetles, with **200** labelled examples.



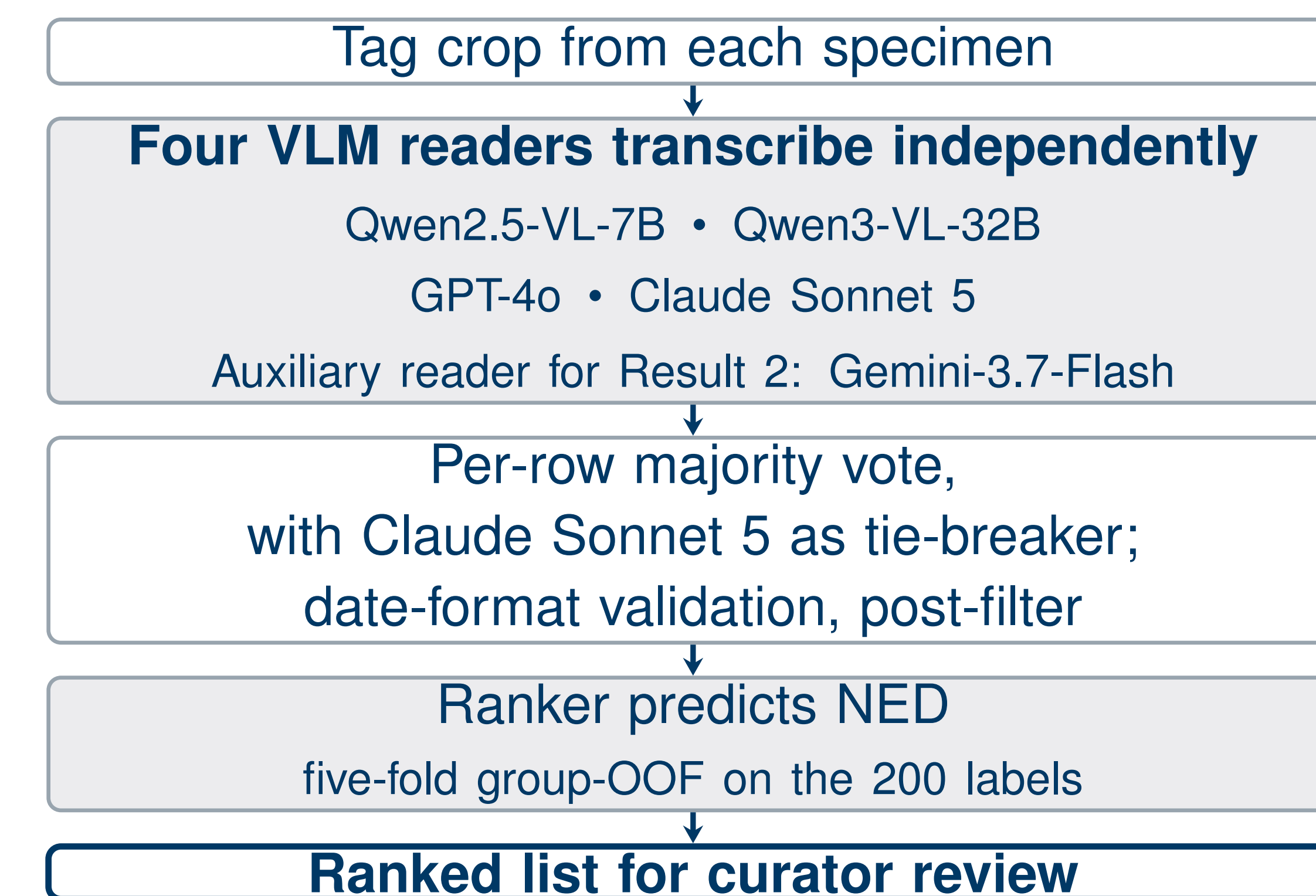
One specimen, two cards: locality/date left, collector right.

What the metric measures

NED = edit distance / max length, case-insensitive; MISSING counts as empty. **AURC** averages cumulative mean NED over every coverage level, then over both fields, the risk-coverage view of selective prediction [4, 5]. **Lower is better**.



Method



All diagnostic analyses use the **200 labelled examples**: locality NED **0.157**, out-of-fold AURC **0.0958**; date NED **0.047**, out-of-fold AURC **0.0347**. The challenge submission also uses *reader substitution* (Result 2).

Two ways four readers disagree

Gold locality: **Veilø Bugt**

Qwen2.5-VL-7B Bjørnø
Qwen3-VL-32B Vilp. Sæg
GPT-4o Viborg
Claude Sonnet 5 Vislp Bugp

Vote output **Vislp | Bugp**
NED 0.40

All four readers produce different locality predictions.

Gold locality: **Grenaa**

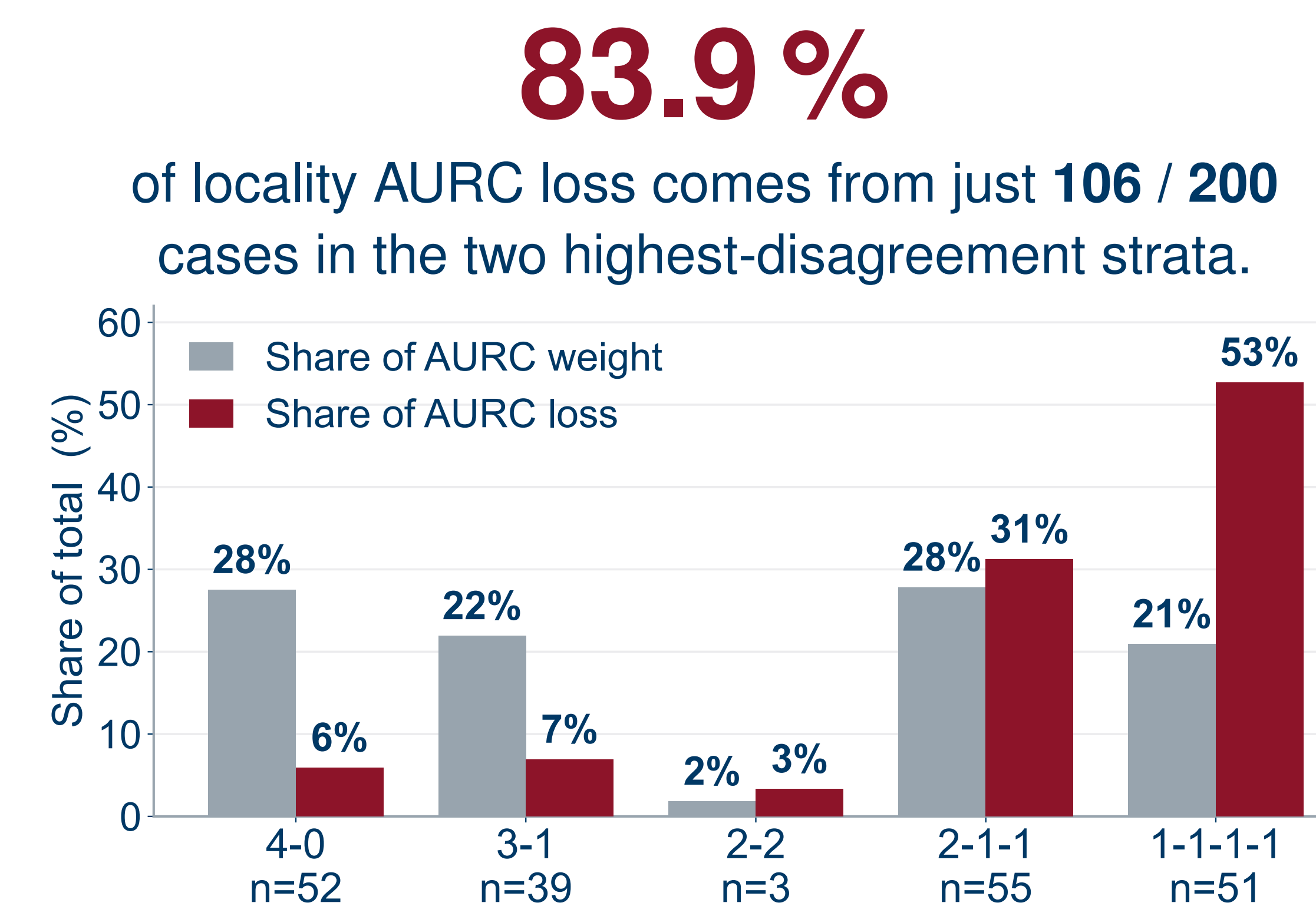
Qwen2.5-VL-7B Grenaa | Brøstrup
Qwen3-VL-32B Grenaa
GPT-4o Grenaa | Brostrup
Claude Sonnet 5 Grenaa Broestrup

Vote output **Grenaa | Broestrup**
NED 0.63

All four transcribe the visible text correctly, but differ in **which label segment they select as locality**.

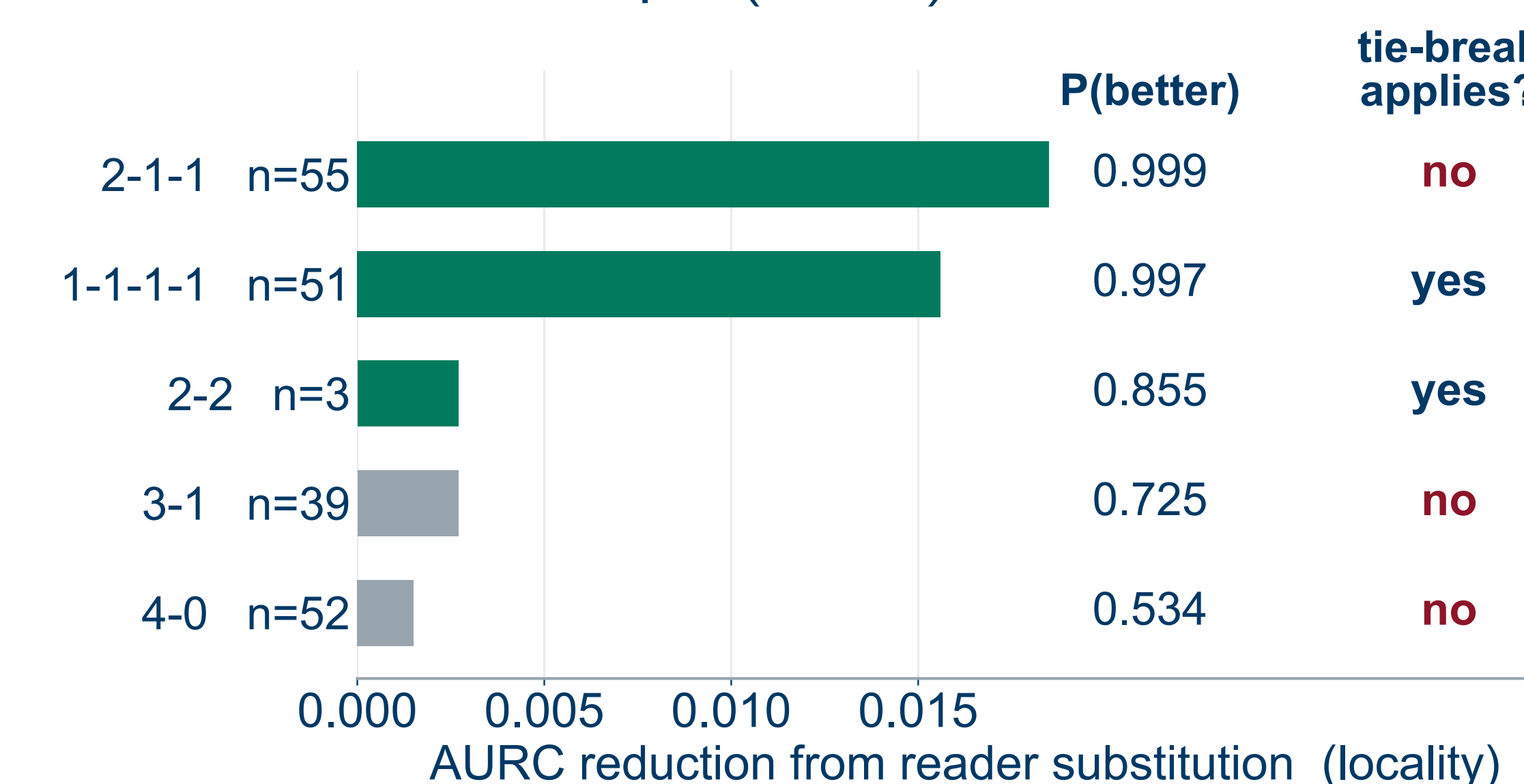
- [1] MuseumSCAT: Specimen Collection Annotation Task @ CVNH ECCV26. Kaggle (2026).
[2] Hedrick, B.P., et al.: Digitization and the Future of Natural History Collections. *BioScience* **70**(3) (2020).
[3] Wiczeorek, J., et al.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* **7**, e29715 (2012).
[4] El-Yaniv, R., Wiener, Y.: On the Foundations of Noise-free Selective Classification. *JMLR* **11**, 1605–1641 (2010).
[5] Geifman, Y., El-Yaniv, R.: Selective Classification for Deep Neural Networks. *NeurIPS* (2017).

Result 1 — AURC loss concentrates in high-disagreement cases



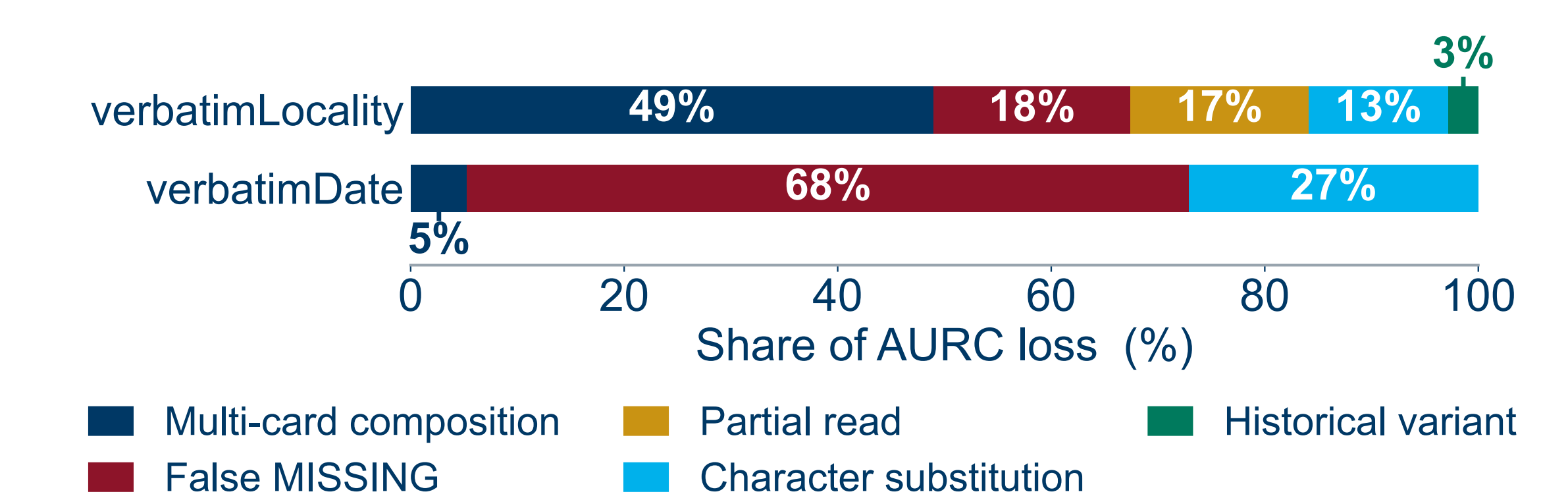
Result 2 — reader substitution improves high-disagreement cases

Gemini-3.7-Flash is the strongest auxiliary reader. Used only as a *tie-breaker*, it can affect 2–2 and 1–1–1–1 cases, but not a 2–1–1 plurality. If it instead *substitutes* for one reader, the largest improvement occurs in the 2–1–1 stratum: AURC decreases by **0.0185**, with bootstrap $P(\text{better}) = 0.999$.



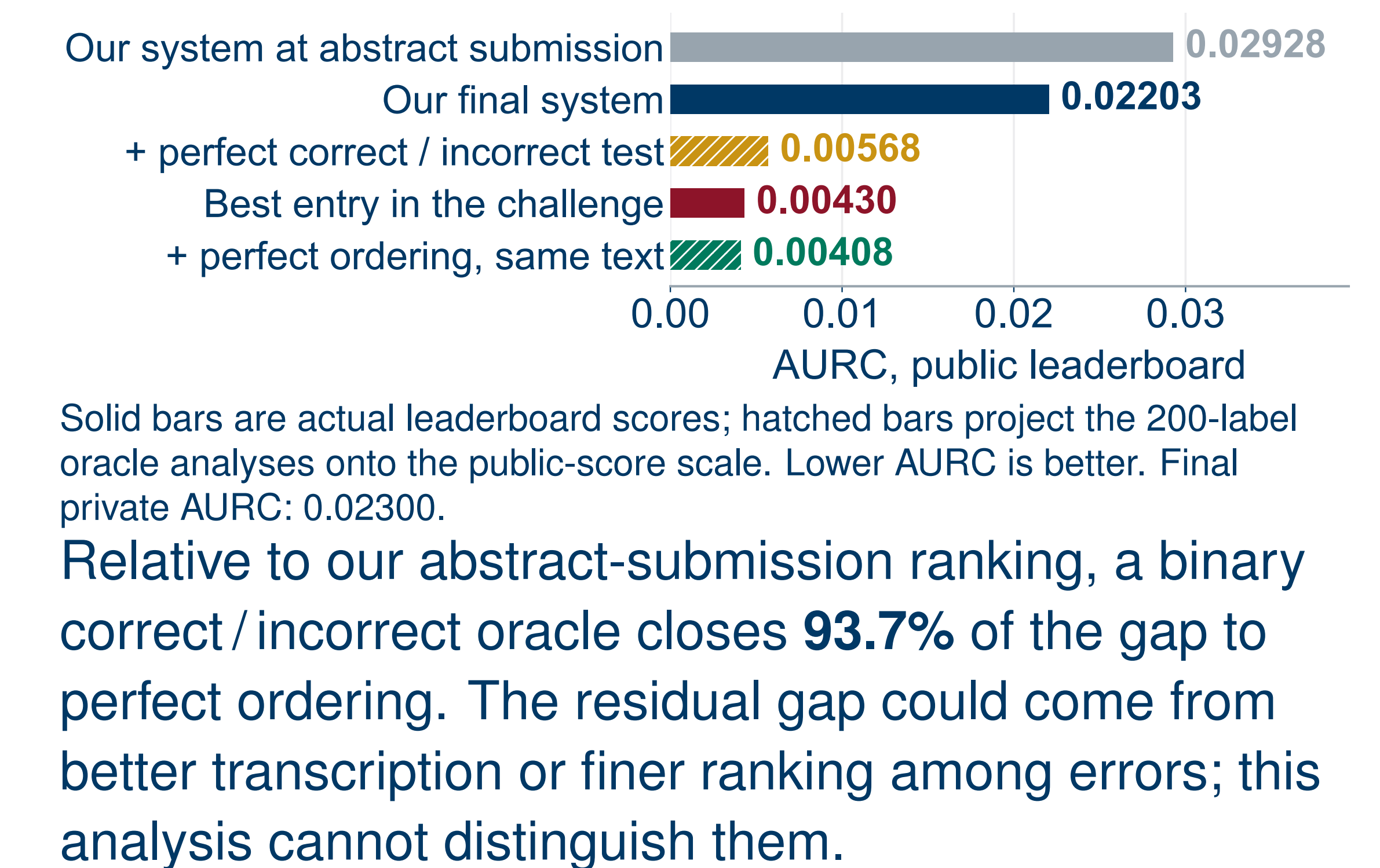
On the **public leaderboard**, targeted reader substitution (**green**) reduced AURC from **0.03336** to **0.02997**; the final submission achieved **0.02928**.

Error composition differs between locality and date



Joining pipe segments changes NED by only **0.0003**: the loss is in *which* segments are emitted.

Most remaining AURC loss comes from ranking



Conclusion

- Reader disagreement** identifies high-risk cases and indicates where reader substitution can improve predictions.
- Reordering the existing predictions recovers **93.7%** of the gap to perfect ordering.
- Further improvement therefore depends primarily on **better confidence ranking**.

Contact



Md Raihan

Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany

✉ md.raihan@fau.de