Episode Book

EPISODE #6584

Why AI Leaderboards Dropped Elo for Bradley-Terry

PUBLISHED	14 November 2026 at 09:22 UTC
DURATION	28:46
LISTEN	myweirdprompts.com/episode/bradley-terry-elo-pairwise-rankings-38383/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	22 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 11:16 UTC

SYNOPSIS

Turning pairwise votes into a reliable ranking requires more than counting wins. Replaying 135,632 real arena votes shows online Elo mostly remembers the last 22 games, shifting one model by 32 places compared to a batch Bradley-Terry fit. This episode unpacks the math behind Bradley-Terry, Elo, and Microsoft's TrueSkill, using a 120,000-vote Indic TTS study and LMArena's switch as real-world benchmarks. It also explains why dropping ties and both-bad votes—over a quarter of all data—artificially inflates leaderboard confidence.

TRANSCRIPT



Corn

Last of three, Herman. Daniel has been walking us through his pairwise evaluation setup — the concept, the tooling, and now the maths. His app already shows Bradley-Terry standings for his voice votes, so today is the episode where we figure out what that number actually is.



Herman

And it's the good part, honestly. Parts one and two were about asking the question and collecting the answers. This is about what happens when a few hundred answers have to become a leaderboard.



Corn

His framing, compressed: he's ranking TTS voice clones by voting on pairs — A better, B better, both bad, about the same. Two candidates, trivial. Five or ten, or two models across hundreds of prompts, and you need a pile of votes turned into a ranking with some confidence attached. He wants the models that turn win-loss records into strengths, he wants to know where ties and both-bad fit, and he wants to know how to choose which pair to show next so he needs fewer votes.



Herman

The good news is this is one of the best-documented corners of applied statistics, and there's a TTS paper from this year that did exactly what he's doing, at a scale that should embarrass all of us — a hundred and twenty thousand pairwise comparisons. So none of this is hypothetical.



Corn

Then let's earn it.



Herman

The core problem first. A pile of pairwise votes is not a ranking. If clone A beat clone B seven times out of ten, and B beat C six out of ten, and C beat A six out of ten — you've got a cycle, and win-loss records don't resolve cycles. What you need is a model that maps each voice onto a strength number such that the observed wins are roughly what those



Herman (continued)

numbers would predict. And then you need to know how much to trust each number.



Herman

It is. There are three families to keep straight, and I'll name them up front so we have a map. Bradley-Terry: a batch fit over all the votes at once — that's the arena standard now. Elo: an online update, one vote at a time — the thing the big arenas moved off of. And the Thurstone family, TrueSkill being the famous one, where every item carries a distribution instead of a point estimate — that's the confidence-native approach. Everything else today is a variation on those three.



Corn

And I'll spoil the ending, because it's the punchline the whole episode earns: the model choice matters less than people think for the order of the ranking, and more than people think for the confidence intervals and the head-to-head odds.



Herman

And the voice-clone angle makes it sharper. The Indic TTS study — Anand and colleagues, seven state-of-the-art TTS systems, ten Indic languages, more than a hundred and twenty thousand pairwise comparisons from nearly two thousand native raters — ran Bradley-Terry on the whole thing and built a multilingual leaderboard. So for once, Daniel's exact use case has a precedent at industrial scale.



Corn

Right. Wine first, though.



Herman

Wine first, because it's the cleanest way in. The canonical Bradley-Terry example is a wine-tasting survey. You can't hand a person forty wines and ask them to produce a ranking — they can't do it. But they can taste two and say which they prefer. Bradley-Terry is the machine that takes all those pairwise judgements and produces the full ranking you never could have asked for directly. Daniel's voice clones are the same animal: nobody can hold ten clones in their head, but anyone can listen to two.



Corn

So how does the machine work?



Herman

It's beautifully small. Every item gets a positive strength number, p. The probability that item i beats item j is p-i over p-i plus p-j. That's the whole forward model. Two equal strengths, fifty-fifty. One twice the other, two-thirds.



Corn

And the useful direction is backwards.



Herman

Right — you don't know the p's, you know the outcomes, and you find the p's that make the outcomes most likely. Maximum likelihood. And here's the cleanest way to say it, the thing I want Daniel to walk away with: Bradley-Terry is logistic regression with one coefficient per item. In ordinary logistic regression you know the parameters and infer the functional form. Here you know the functional form — it's that little fraction — and you infer the parameters.



Corn

One coefficient per voice clone. That's the whole leaderboard.



Herman

Fitting it is an iteration Zermelo worked out in the nineteen-twenties — Bradley and Terry just presented it in nineteen fifty-two — and each round you normalize by the geometric mean, because the strengths are only identified up to a multiplicative constant. Double every p and every prediction is unchanged. So you pin the scale, or you report only differences, which is what everyone actually does.



Corn

Which is why a bare arena score of, say, one thousand four hundred and fifty means nothing.



Herman

Nothing at all. Only differences mean anything. Thirty points above model X is an expected win rate of roughly fifty-four percent against it. Same number, now it's a sentence.



Corn

Now Elo. Everyone knows Elo from chess, and you're about to tell me it's the same model wearing a hat.



Herman

Elo is Bradley-Terry with a scale factor of four hundred and base ten — the probability A beats B is one over one plus ten to the minus d over four hundred, where d is the rating gap. A hundred-point gap gives you a sixty-four percent favourite. Four hundred points, ninety-one percent.



Corn

So if the maths is identical, why did the arenas move off it? That's the story I came for.



Herman

The difference is operational, and it's brutal. Elo is an online update: after each vote, the winner's rating moves up by K times the result minus the expected result, and the loser moves the same amount down. Bradley-Terry is a batch fit: all votes, at once, order irrelevant. And it turns out order matters enormously under Elo, because a running rating is not a summary of all its votes.



Corn

It's a memory of the recent ones.



Herman

Quantifiably. A rating's error shrinks by a fixed fraction per vote, so at K equals four, a rating mostly reflects its last hundred and seventy-four votes. At K equals thirty-two, its last twenty-two.



Corn

Twenty-two. You could rank a chess club on a good weekend.



Herman

And here's the scandal. Someone took a hundred and thirty-five thousand six hundred and thirty-two real arena votes across fifty-two models and replayed them — same votes, in the order they were cast — with online Elo at K equals thirty-two. Claude Sonnet four came sixth of fifty-two. Then they fit all the votes at once with Bradley-Terry. Thirty-eighth. Thirty-two places, from identical data.



Corn

Thirty-two places. That's not noise, that's a different leaderboard.



Herman

Shuffling the vote order alone moved a model's rating by up to plus or minus thirty-four points at K equals four, while the Bradley-Terry fit on the same data has a median ninety-five percent interval of plus or minus seven point seven. And when LMArena switched, in December twenty twenty-three, their stated reason was almost embarrassingly practical. They said, we have access to the entire history of all games for all models, and so we don't need a decentralized algorithm — and most models are static. They'd been averaging Elo over a thousand shuffles of the votes to patch the order problem, and they conceded the intervals came out too wide.



Corn

Too wide. Their patch was making honest numbers look evasive.



Herman

The analogy that makes it stick is a chess club. Elo is a ladder — recent games count most, which is right for a junior who's actively improving. Bradley-Terry is a season table — every player's strength treated as fixed for the season, so the order the games happened in doesn't matter. A model checkpoint behind an arena is a fixed player, not a growing junior.



Corn

That's the part people get wrong, then — they assume a running rating summarizes the votes. It summarizes the last twenty-two of them.



Herman

And more votes never fix it, because the short memory is a noise floor. Now the third family. TrueSkill, which Microsoft built for Xbox matchmaking. The structural difference is that TrueSkill carries a distribution per player — a mean and an uncertainty — instead of a point estimate. Which is literally what Daniel asked for when he said confidence attached. Plain Bradley-Terry gives you a point estimate and you bolt the uncertainty on afterwards, through standard errors or bootstrap resampling. TrueSkill has the uncertainty inside the model. It's a Thurstonian model — Gaussian score distributions under the hood — and it handles more than two players at once, which pairs can't.



Corn

Which raises the obvious question for Daniel: if he's only ever comparing pairs, is the Plackett-Luce generalization — the one that handles ranking k items — worth anything to him?



Herman

Directly, no — for any two items, Plackett-Luce reduces to Bradley-Terry. But hold that thought for the last third of the episode, because the generalization is where the real savings live, and it's not where people expect.



Corn

He said, loading the gun.



Herman

Loading it carefully. To answer Daniel's first question head on: how do you turn a pile of pairwise votes into a ranking with confidence attached? Two routes. Fit Bradley-Terry by maximum likelihood and get your uncertainty from standard errors or from bootstrap resampling — the VidyalAI analysis uses a thousand resamples, and LMArena's leaderboard reports bootstrap ninety-five percent intervals as ranges rather than point estimates. Or go TrueSkill and carry the distribution natively. Both are defensible. Daniel's app already does the first one, and the first one is the arena standard, so he's in good company.



Corn

Good. Now the part his interface makes unavoidable. Not every vote is A better or B better.



Herman

Not even most of them, in practice. In that same hundred and thirty-five thousand vote dataset — mid-April to late July last year — fifteen point nine percent were ties and eleven point six percent were both bad. Together, twenty-seven and a half percent of every vote cast.



Corn

More than a quarter of the data sitting in the two bins people are tempted to throw away.



Herman

And the temptation is wrong, in a measurable way. The simple convention — and this is what both Elo and the Bradley-Terry fit in that analysis actually do — is that a tie counts as half a win and half a loss. You record a one if the rater prefers A, a half for a tie, a zero if they prefer B. And that is why the simple convention treats a tie as half a win and half a loss for each model.



Hilbert

Hilbert's Mid-Show Haiku. Ties. Half a win each, and nobody is ranked. The confidence interval is wider than the gap.



Corn

Every time, on a slow news day.



Herman

So what happens if you drop them instead? Someone checked. Dropping ties and both bad barely reorders the table — at most two places. But it stretches it: the standard deviation of the ratings grows from fifty-seven and a half points to nearly eighty. Because the votes you kept look more decisive. The ranking looks sharper than the data supports.



Corn

The leaderboard gets more confident as you feed it less evidence. That's a machine doing the opposite of its job.



Herman

And it's not just stability — ties are where the modelling gains actually come from. There's a paper on one point eight million arena votes showing a four-to-six parameter correction predicts held-out votes better than Bradley-Terry and better than models with hundreds of parameters, mostly by modelling ties. All that machinery, and the payoff was in the bin everyone was discarding.



Corn

Does the fancy tie treatment change the answer, though?



Herman

Mostly no, which is the comfortable result. Rao-Kupper, from nineteen sixty-seven, gives ties their own parameter instead of forcing the half-win convention — on the real arena data it fitted a tie parameter of one point eight two and moved no model more than two



Herman (continued)

places. So the half-win convention is a safe simplification. If Daniel wants to go further there's the factored tie model from ICLR last year, which models ties explicitly and adds covariance between competitors so you get intuitive groupings into performance tiers — but that's polish, not necessity.



Corn

Hold on. Before we file ties as solved — they're solved as a scoring problem. Is "both bad" solved at all?



Herman

No. And this is a genuine gap, and I want to flag it properly because Daniel's interface has both buttons and the literature mostly has one. A tie says these two are equal. Both bad says these two both failed an acceptability bar — which is a statement about the pair against a standard, not about the pair against each other. A Bradley-Terry strength parameter has no slot for that. You'd need an extra term, some kind of quality floor per item, and I could not find a standard model for it. The treatments that exist — Rao-Kupper, the factored tie model, partial rankings via Bayesian clustering — all handle ties and treat both bad as either a tie or a discard.



Corn

So a quarter of the votes, and one of the four buttons, and the entire literature shrugs.



Herman

The best anyone says is the pragmatic line — a forced choice between two equally broken answers is a coin flip that looks like data. Which is an argument for collecting the button, not a model of what the button means.



Corn

My instinct — and tell me if the maths backs this — is that both bad is actually the more informative of the two odd buttons. If I'm choosing between two voice clones and one of them hallucinated a word, both bad is the vote that tells you the clone is unusable regardless of how the other one sounds.



Herman

The maths backs the instinct, in an oblique way. There's work on partial rankings arguing most inference-based methods choose to assign each item to a unique rank or score, suggesting a meaningful distinction when there is none — and that the honest output is a ranking with ties in it, distinguishing items only when the data supports it. Both bad is the same idea from below: some items shouldn't be on the ladder at all. Relatedly, people have embedded Bradley-Terry in a stochastic block model so items cluster and share a tied rank — applied to twenty-odd years of ATP tennis, the top hundred players partition into three or four tiers in most seasons. Which is probably the truthful shape of Daniel's leaderboard too. Not one through ten. Tier, tier, tier.



Corn

He'll like that. It matches the interface — "about the same" is a tier signal, not an apology.



Herman

Now his last question, and the one with the most folklore attached: choosing which pair to show next, so you need fewer votes. The framing in the literature is active learning from noisy pairwise comparisons, and there's real work here — robust elicitation procedures, utility-based sampling that picks the query expected to most improve the final recommendation. And it does work on actual humans, not just simulation: they deployed one of these robust procedures on about two hundred Mechanical Turk workers for COVID patient prioritization, and it beat asking random queries by twenty-one percent.



Corn

Twenty-one percent against random. Write the caveat on the same line as the number.



Herman

It has to be. Random is the worst baseline that still gets published. Nobody in these results claims active selection beats a well-designed balanced schedule by an order of magnitude. And there's a hard theoretical ceiling on top of that, which I find clarifying. Saha and Gopalan proved a sample complexity bound for adaptively ranking n items — and proved the matching lower bound, which shows there is essentially no improvement possible from the pairwise setting. Cleverness in choosing pairs cannot beat the pairwise limit. What does buy you something is asking a richer question: elicit top-m feedback from a subset — rank these three, best of these five — and you get an m-fold reduction in sample complexity, and that's order-wise unimprovable too.



Corn

So the answer to "which pair do I show next" is: it barely matters, and the win is in asking something bigger than a pair.



Herman

That's the provable answer, yes. For Daniel specifically, the practical translation is: with five or ten clones, a balanced round-robin schedule with randomized left-right order will get him essentially everything adaptivity would, and the cost of the randomness is worth paying twice over — there's a result on converting systematic position bias into zero-mean noise by randomizing direction, so randomize which clip plays first and any left-side bias becomes noise the fit averages away instead of bias it bakes in.



Corn

He's comparing two models across hundreds of prompts, though. Not ten clones once — one pair, many prompts.



Herman

Which changes the unit of analysis, and it's where the vote-count question bites. The basic arithmetic: a win rate estimated from n comparisons has standard error the square root of theta times one minus theta over n. A hundred comparisons near fifty-fifty gives you a five-point standard error. Telling a fifty-five percent win rate from fifty-fifty, with eighty percent power at the five percent level, takes about eight hundred comparisons.



Corn

Eight hundred for one pair. His ears are going to file a complaint.



Herman

And it gets worse at arena scale — separating two models ten points apart in a real arena took about thirteen thousand seven hundred votes each. That's an LLM-arena figure, not speech, and nobody has published the TTS equivalent — but the shape transfers, and it's the single number I'd want him to keep: distinguishing near-equals is brutally expensive, and most of his pairs will be near-equals.



Corn

Which is the argument for the tier answer, again. Don't pay thirteen thousand votes to separate two things a coin flip separates anyway.



Herman

And here's the subtlety about the confidence once he has it. Overlapping intervals are not a tie. WebDev Arena in October had Claude Opus at eighteen twenty from under nineteen hundred votes, interval eighteen-oh-three to eighteen thirty-seven, and GPT-6 Astra at eighteen hundred, interval seventeen eighty-four to eighteen seventeen. The intervals overlap. And that overlap does not settle the question, because checking whether two intervals overlap is a much stricter test than it looks — the correct test is a direct comparison of the two, not eyeballing error bars on a chart.



Corn

Every reader of every leaderboard chart is doing the wrong test.



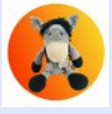
Herman

Constantly. Now the caveat I've been saving, and it's the one that applies to Daniel more than to anyone, because he's ranking near-copies. Bradley-Terry assumes Luce's choice axiom, and that assumption measurably breaks for near-copies. On one point eight million arena votes: comparisons between versions of one model were one point three seven times as decisive as Bradley-Terry predicted. Between sizes of one generation, one point two six. Between models of the same lab, one point one two.



Corn

Two checkpoints of the same TTS model — which is literally what a voice clone ladder is.



Herman

That's the example. The leaderboard order barely moves — the ranking survives — but the head-to-head odds are systematically wrong. The model will tell you clone A beats clone B sixty-four percent of the time when raters actually said so more decisively than that. Raters separate near-copies more sharply than their strengths imply, and how alike the two answers are worded and formatted tracks the effect. For voice clones, how alike they sound.



Corn

So his standings table will be roughly right and his odds will be overconfident. Which, given the episode's punchline, is exactly where we said the model choice would show up.



Herman

And the TTS-specific evidence says he can do this at all, which is worth dwelling on for a minute. The Indic study's whole motivation was the failure of MOS — mean opinion score, the one-to-five absolute rating that speech has used forever. Their line is that MOS relies on absolute ratings sensitive to individual rater calibration — some raters hand out fours, some hand out twos, and the scale never agrees with itself — while pairwise preference captures direct comparisons and enables statistically grounded rankings. And there's a rethinking-MUSHRA paper making the adjacent point: MOS tests fail to differentiate between similar models, and MUSHRA's insistence on a human reference unduly penalises modern systems that can exceed human speech quality.



Corn

The reference ceiling. When the students start beating the teacher, grading against the teacher stops making sense.



Herman

If he's tempted to automate — use a quality predictor instead of his own ears — the finding there is sharp. UTMOS, DNSMOS, SCOREQ, tested on pairwise tasks: when one clip carries audible defects, the predictors agree with listeners. Once both clips are clean, no single predictor reliably identifies the preferred sample, and several fall below the accuracy of simply picking the longest clip.



Corn

Longer always wins. That's not a judge, that's a tape measure.



Herman

It's the strongest argument for keeping humans in the loop on voice clones specifically, at least at the top of the quality range, where Daniel's clones will live. The design detail worth stealing from the Indic study: they collected six perceptual dimensions per comparison — intelligibility, expressiveness, voice quality, liveliness, noise, hallucinations — in addition to overall preference, because overall preference alone provides limited insight into why.



Herman (continued)

One vote plus a couple of tags. And they released the data publicly.



Corn

We did an episode on the metric side of this — benchmarking custom ASR tools, beyond the word error rate — and this is the same lesson arriving from the other direction. The number you can defend beats the number that's easy to collect.



Herman

One more TTS result, because it validates the both-bad button by accident. A voice-editing model's human evaluation — four hundred judgments on a forty-sample subset — found fifty-seven point two percent of listeners rated the edited speech as perceptually equal to the original, twenty-four point eight preferred the original, eighteen percent preferred the edit.



Corn

Fifty-seven percent "about the same.



Herman

On close pairs, that's what the distribution looks like. Which is why the standard deviation stretches when you throw those votes out — you're not removing noise, you're removing the evidence that the race is close. There's even a convergence result showing pairwise comparisons need more of them when ties are frequent. The ties aren't just harmless. They're the expensive part.



Corn

Nobody's modeled both bad. Herman, I think Daniel found the gap.



Hilbert

Fifty-seven point two percent of four hundred. The forty was the subset of RealEdit, and four hundred was the judgments, ten to a sample, so if you're rounding to fifty-seven you've thrown away the two-tenths that did the work. Herman had it right down to the decimal and then Corn said fifty-seven flat.



Corn

I said fifty-seven.



Hilbert

You did. I've been voting on pairs myself since September. Two recordings of the same book, one from the nineties with a human narrator, one of these new synthetic ones. Eleven chapters so far, each pair scored the same way Daniel does. Better, worse, or the same. All on paper, in pencil — pen implies you've finished deciding. The same column is nine of the eleven chapters.



Herman

Nine of eleven ties. What's the split on the other two?



Hilbert

The machine reads the numbers and the street names better. The man reads everything else. Chapter nine I called a tie twice because the first time I'd left the window open and the traffic had opinions. I wrote to the council about it, actually — not the window, the street pronunciations, whether the local recordings office had a canonical list for the narrator's street names. A man named Dror in the records department has been finding out for me since October. He phones on Tuesdays with progress. The bank also declined the audiobook purchase three times, which cost me an afternoon, but the votes themselves cost nothing. I stop at ten comparisons per pair. Eleven if it's five and five. That rule is mine and it has not failed me.



Corn

Ten and stop. You invented a fixed budget per pair.



Hilbert

I invented nothing. I got tired after ten. There's a difference, Corn, and Herman of all people knows I don't invent things at this desk. Chapter twelve is queued for tonight. I expect another tie.



Herman

Send us the final table when the book's done.



Corn

The ranking is the easy part. That's what the whole episode keeps running into — the order will come out of Daniel's app almost regardless of what he does. The confidence intervals, the tie rates, the head-to-head odds on near-identical clones, that's where the model choice actually shows up, and that's where the open questions live too. No standard model for both bad, no published active-selection work on listening tests, and the thirteen-thousand-votes figure is borrowed from text arenas, not speech.



Herman

Though a fixed ten-vote budget per pair, arrived at by getting tired, is more or less where the theory lands anyway — the pairwise limit means the balance of the effort belongs in asking better questions, not in agonizing over which pair is next. If Daniel ever adds a third button that plays three clips and asks which is best, the maths says that's the real discount.



Corn

So: batch-fit Bradley-Terry, keep every tie, randomize the left clip, report intervals as ranges, and don't trust the odds between two checkpoints of the same model. Thanks to Hilbert Flumingtop for producing this one, and the whole three-part run. If you enjoyed this one, try episode eleven, How Does Fine Tuning Work Anyway; episode nine, Benchmarking Custom ASR Tools - Beyond The WER; and episode fifteen, AI Gets Personal. This has been My Weird Prompts. If the pairwise series earned its keep, a review on Spotify is the cheapest vote you'll ever cast, and unlike a tie it counts whole. We'll be back soon.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Anand, Sankar, Sethi et al., Preferences of a Voice-First Nation: Large-Scale Pairwise Evaluation and Preference Analysis for TTS in Indian Languages, arXiv:2604.21481v2, revised 2026-06-23 [primary]**

   https://arxiv.org/abs/2604.21481v2

2. **Wikipedia, Bradley**
   Terry model, last edited 2026-09-28, (retrieved 2026-11-14)
   https://en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model

3. **Wikipedia, Elo rating system, last edited 2026-11-12**
   (retrieved 2026-11-14)
   https://en.wikipedia.org/wiki/Elo_rating_system

4. **Wikipedia, TrueSkill, last edited 2026-10-06**
   (retrieved 2026-11-14)
   https://en.wikipedia.org/wiki/TrueSkill

5. **VidyalAI, Arenas and Elo Rankings: How Pairwise Votes Become a Leaderboard (Part 7 of 11)**
   (retrieved 2026-11-14)
   https://vidyal.ai/blog/chatbot-arena-elo/

6. **Zero to Frontier, 22.7 Human evaluation and arenas, Chapter 22**
   (retrieved 2026-11-14)
   https://zerotofrontier.app/evaluation/arenas

7. **Marovi AI, LMSYS Arena**
   (retrieved 2026-11-14)
   https://marovi.ai/LMSYS_Arena

8. **AI Zero to Hero, Human-Preference Arenas and LLM Judges**
   (retrieved 2026-11-14)
   https://ai.noblerworks.com/explainers/benchmarks/human-preference-arenas/

9. **Ameli, Zhuang, Stoica et al., A Statistical Framework for Ranking LLM-Based Chatbots, arXiv:2412.18407v2, ICLR 2025**

   https://arxiv.org/abs/2412.18407v2

10. **Bakshi, Bose, Bose et al., Similar Models, Broken Odds: Where Luce's Axiom Fails in Chatbot Arena, arXiv:2610.19095v2, revised 2026-10-18**

    https://arxiv.org/abs/2610.19095v2

11. **Morel-Balbi & Kirkley, Estimation of partial rankings from sparse, noisy comparisons, arXiv:2501.02505v4, revised 2026-01-08**

    https://arxiv.org/abs/2501.02505v4

12. **Santi & Friel, The Bradley-Terry Stochastic Block Model, arXiv:2511.03467v1, 2025-11-05**
    https://arxiv.org/abs/2511.03467v1

13. **Levtsov & Ustalov, Confidence and Stability of Global and Pairwise Scores in NLP Evaluation, arXiv:2507.01633v1, ACL SRW 2025**

https://arxiv.org/abs/2507.01633v1

14. **Paschmann, Kaplan, Nattero et al., Active Learners as Efficient PRP Rerankers, arXiv:2605.14236v2, revised 2026-05-15**

https://arxiv.org/abs/2605.14236v2

15. **Boroomand & Wright, Personalized Recommendations via Active Utility-based Pairwise Sampling, arXiv:2508.14911v1, 2025-08-12**

https://arxiv.org/abs/2508.14911v1

16. **Vayanos, Ye, McElfresh et al., Robust Active Preference Elicitation, arXiv:2003.01899v2, revised 2021-12-08**

https://arxiv.org/abs/2003.01899v2

17. **Johnston, Vossler, Blessenohl et al., Deploying a Robust Active Preference Elicitation Algorithm on MTurk, arXiv:2306.04061v2, EAAMO 2023**

https://arxiv.org/abs/2306.04061v2

18. **Saha & Gopalan, Active Ranking with Subset-wise Preferences, arXiv:1810.10321v2, AISTATS 2019**

https://arxiv.org/abs/1810.10321v2

19. **Varadhan, Gulati, Sankar et al., Rethinking MUSHRA: Addressing Modern Challenges in Text-to-Speech Evaluation, arXiv:2411.12719v3, TMLR**

https://arxiv.org/abs/2411.12719v3

20. **Asonitis, Zuluaga Gomez, Verdini et al., The Limits of Reference-Free Speech Quality Metrics as Evaluators and Rewards on Modern Text-to-Speech, arXiv:2609.13150v1, 2026-07-02**

https://arxiv.org/abs/2609.13150v1

21. **Mohammad, Zhussip, Lefkimmiatis, Speak, Edit, Repeat: High-Fidelity Voice Editing and Zero-Shot TTS with Cross-Attentive Mamba, arXiv:2510.04738v1, 2025-10-06**

https://arxiv.org/abs/2510.04738v1

22. **Landesberg, When LLM Judge Scores Look Good but Best-of-N Decisions Fail, arXiv:2603.12520v1, 2026-03-12**

https://arxiv.org/abs/2603.12520v1