12^{th} International Conference on Computational Social Science IC^2S^2
July 28-31, 2026, Burlington, VT, United States

Beyond Single Ground Truth: Reference Monism as Epistemic Injustice in ASR Evaluation

Keywords: ASR, epistemic injustice, evaluation, ground truth, AI bias

Extended Abstract

Automatic speech recognition (ASR) systems are evaluated by comparing system output to a ground truth transcript, with Word Error Rate (WER) quantifying the distance between them. This methodology rests on a false assumption so naturalized it typically goes unstated: that there exists a single correct transcription of a speech event against which all outputs should be measured. Ground truth transcripts are not discovered but constructed by human annotators following conventions that encode normative assumptions about which features of speech matter. A verbatim convention preserves fillers ("um," "uh"), false starts, repetitions, and repairs. A non-verbatim convention removes these, producing clean readable text. A legal convention strictly preserves hedges, false starts, and qualifications relevant for evidentiary purposes, but handles standard speech disfluencies differently than a purely linguistic verbatim transcript. Figure 1 illustrates this directly, showing how a single utterance from a speaker with aphasia yields three legitimate transcriptions that would produce substantially different WER scores for the same system output.

We argue that **reference monism (Figure 1)** – the enforcement of a single transcription convention as the sole ground truth – commits epistemic injustice in Miranda Fricker's sense [1]. The harm is not merely differential performance. It is that ***the evaluative infrastructure itself lacks the interpretive resources to recognize certain speakers' communicative contributions as legitimate.*** Speakers with aphasia, whose speech is characterized by clinically meaningful disfluencies, are the central case: they are systematically penalized when evaluated against "clean" references that treat those disfluencies as errors, even though verbatim clinical conventions – developed precisely for this population – would render the same output far more favorably. The injustice is structural and avoidable. We synthesize philosophical tradition [1, 3, 5, 2] and introduce the **hermeneutical gap**: the distance between a speaker's communicative contribution and the interpretive resources available in a transcription convention to render that contribution intelligible. The hermeneutical gap is not a property of speakers but of evaluative frameworks – it measures not how "clearly" someone speaks but how well the interpretive infrastructure accommodates their communicative practices.

We then formalize these philosophical concepts into empirically tractable quantities. We define **Epistemic Injustice Distance (EID)** for a speaker group as the additional evaluation penalty incurred because a dominant convention is enforced, even when other legitimate conventions in the policy set would judge the same system output more favorably. We define **ΔEID** as the comparative injustice between groups – the structural evaluation burden borne by one group that another does not bear. Crucially, EID is non-negative by construction and avoidable by construction: it is zero if evaluation accepts any legitimate reference, which means the burden exists only because institutions enforce a single framework when multiple legitimate ones are available.

We provide empirical evidence using the AphasiaBank corpus [4] (a stratified test set of approximately 9 hours, 29 control and 30 aphasic speakers). We evaluate seven ASR systems

– including Rev AI[1] v2 and v3 (verbatim, non-verbatim, and legal modes), Whisper-large-v3[2] and CrisperWhisper[3] – against all three transcription conventions simultaneously, each produced by professional annotators following established guidelines.

The results are stark. For Rev AI v2 in verbatim mode, WER ranges from 9.81% to 17.38% depending solely on which convention defines ground truth – a 1.8× difference for identical system output (Table 1). This is not a marginal effect: it is larger than the performance differences typically used to rank systems on leaderboards. Crucially, this convention-dependence is not uniform across speaker groups. Under non-verbatim enforcement, the fairness gap between control and non-fluent aphasic speakers is 21.68 percentage points; under verbatim enforcement, it narrows to 13.50 pp – a 60% change in apparent fairness arising purely from convention choice, not system behavior (Table 2). Which system appears "moderately unfair" versus "severely unfair" depends entirely on which interpretive standard is enforced.

EID and ΔEID calculations confirm the philosophical prediction: enforcing non-verbatim conventions as the default imposes 8.18 additional percentage points of structural burden on non-fluent aphasic speakers compared to control speakers with Rev AI v2 verbatim – nearly three times the burden borne by controls. Inter-reference distance analysis further quantifies the hermeneutical gap directly: verbatim and non-verbatim references diverge by approximately 7–11% WER across anchoring systems (Table 4), representing an irreducible penalty for any speaker whose natural communicative patterns align with verbatim conventions but who is evaluated against non-verbatim standards. Even a perfect ASR system cannot fall below this distance.

Our practical recommendation is **WER-Range**: ***rather than reporting a single WER number, report the range of performance values across legitimate conventions.*** Instead of "Rev AI v2 achieves 9.81% WER on AphasiaBank," report "Rev AI v2 achieves WER-Range [9.81%, 17.38%] across verbatim, non-verbatim, and legal conventions." WER-Range disaggregated by speaker group reveals that non-fluent aphasic speakers face a range width of 12.91 pp – nearly three times the 4.73 pp range of control speakers – exposing differential vulnerability to convention choice that single-number reporting renders invisible.

We address the cost objection directly. Multi-reference annotation is evaluation infrastructure incurred once per benchmark and reused across all subsequent system assessments. Computational costs (inference, WER calculation) are invariant to reference multiplicity. Where human annotation costs are genuinely prohibitive, we propose staged implementation: convention-labeled reporting first (zero cost, "WER under non-verbatim convention" rather than just "WER"), diagnostic multi-reference evaluation for representative subsamples second, and strategic plural ground truth for populations where convention-dependence is established third. Even when constraints permit only one convention, labeling it as such costs nothing and changes everything – it distinguishes epistemic humility ("we report WER under convention X while acknowledging this limitation") from epistemic monism ("we report WER as if X represents the true transcription").

More broadly, this work locates a form of structural disparity that standard group fairness metrics are architecturally unable to detect. Standard metrics presuppose reference monism: they study prediction distributions conditional on a fixed ground truth. Our approach operates upstream, asking how the choice of ground truth itself structures the fairness assessments conducted downstream. EID and ΔEID complement rather than replace existing fairness metrics; they diagnose when apparent "model bias" is in fact an artifact of evaluative infrastructure.

[1] `https://www.rev.ai/`

[2] `https://huggingface.co/openai/whisper-large-v3`

[3] `https://huggingface.co/nyrahealth/CrisperWhisper`

References

[1] Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing*. Oxford university press, 2007.

[2] Hans-Georg Gadamer. *Truth and method*. A&C Black, 2013.

[3] Trystan S Goetze. Hermeneutical dissent and the species of hermeneutical injustice. *Hypatia*, 33(1):73–90, 2018.

[4] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25:1286–1307, 2011.

[5] Gaile Pohlhaus Jr. Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance. *Hypatia*, 27(4):715–735, 2012.

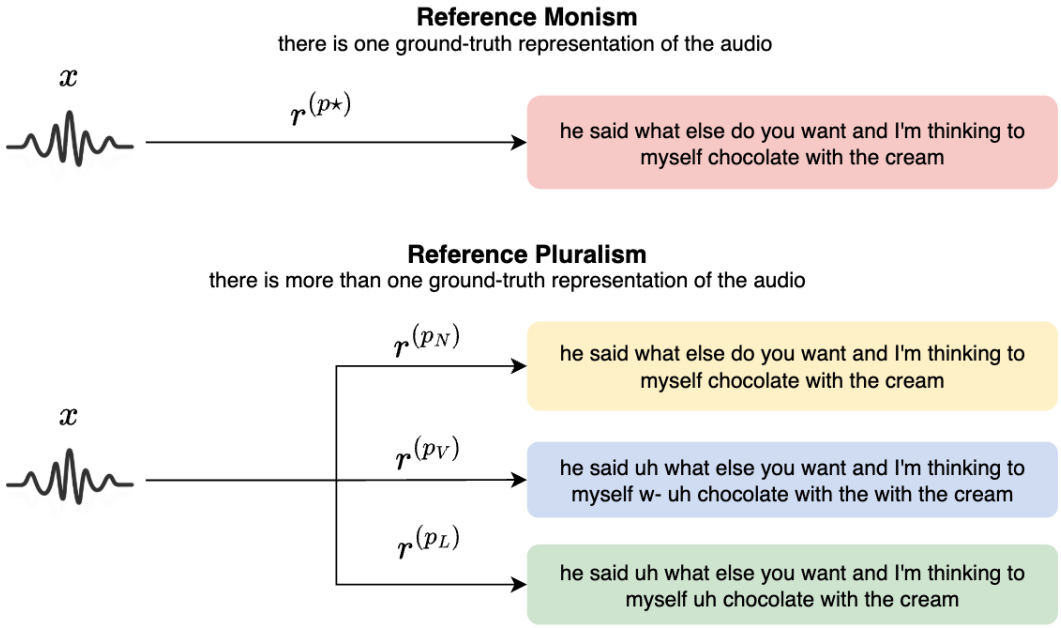

Figure 1: Reference monism enforces a single transcription convention as ground truth, while reference pluralism recognizes multiple legitimate interpretations of the same utterance.

Table 1: WER (%) by ASR system and reference convention (overall test set). Bold indicates lowest WER for each system. Systems perform best when evaluated against "matching" conventions – verbatim systems against verbatim references, etc.

ASR System	**Verbatim** p_V	**Non-verbatim** p_N	**Legal** p_L
Rev AI v2 (verbatim)	**9.81**	17.38	10.46
Rev AI v2 (non-verbatim)	16.18	**9.04**	11.46
Rev AI v3 (verbatim)	**10.60**	17.83	11.94
Rev AI v3 (non-verbatim)	16.95	**9.60**	12.71
Rev AI v3 (legal)	16.00	14.76	**10.96**
Whisper-large-v3	23.85	**19.19**	20.48
CrisperWhisper	**29.67**	30.65	26.20

Table 2: WER (%) by speaker group and reference convention

ASR System	Group	Verbatim	Non-verbatim	Legal
Rev AI v2 (verbatim)	Control	4.03	8.76	5.71
	Fluent aphasia	14.95	26.06	13.92
	Non-fluent aphasia	17.53	30.44	19.60
Rev AI v3 (verbatim)	Control	7.76	11.65	9.30
	Fluent aphasia	18.85	28.72	18.72
	Non-fluent aphasia	24.51	32.40	26.43
Rev AI v3 (non-verbatim)	Control	11.75	7.77	9.32
	Fluent aphasia	26.26	17.05	21.68
	Non-fluent aphasia	35.35	21.61	26.06

Table 3: Edit operation breakdown (%) for Rev AI v2 (verbatim) across reference conventions

Reference	WER	Insertions	Deletions	Substitutions
Verbatim (p_V)	9.81%	2.18%	2.81%	4.82%
Non-verbatim (p_N)	17.38%	12.26%	1.30%	3.82%
Legal (p_L)	10.46%	4.84%	2.33%	3.30%

Table 4: WER-Range by ASR system (top) and by speaker group for Rev AI v2 verbatim (bottom). Range width measures vulnerability to convention choice; wider ranges indicate greater sensitivity to which standard defines ground truth.

ASR System	WER-Range	Range Width
Rev AI v2 (verbatim)	[9.81%, 17.38%]	7.57 pp
Rev AI v2 (non-verbatim)	[9.04%, 16.18%]	7.14 pp
Rev AI v3 (verbatim)	[10.60%, 17.83%]	7.23 pp
Rev AI v3 (non-verbatim)	[9.60%, 16.95%]	7.35 pp
Rev AI v3 (legal)	[10.96%, 16.00%]	5.04 pp
Whisper-large-v3	[19.19%, 23.85%]	4.66 pp
CrisperWhisper	[26.20%, 30.65%]	4.45 pp
Speaker Group	**WER-Range**	**Range Width**
Control	[4.03%, 8.76%]	4.73 pp
Fluent aphasia	[13.92%, 26.06%]	12.14 pp
Non-fluent aphasia	[17.53%, 30.44%]	12.91 pp