From Output Metrics to Cognitive Process: Disfluency and Lexical Entrainment for Evaluating Real-Time AI-Mediated Communication

Anonymous ACL submission

Abstract

As NLP moves from post-hoc text processing into real-time AI-mediated communication, evaluation needs to follow. We argue that a new mission for NLP evaluation is to instrument the human production and comprehension pipelines that AI now intervenes in, not only the AI's outputs. Reference-based metrics (WER, BLEU, task accuracy) cannot capture what happens inside the speaker's production pipeline when an AI intervenes mid-turn, nor what is lost on the listener's side when AI "smooths" the partner's speech. This position paper argues that *speech disfluency* and *lexical entrainment* together provide a theoretically grounded, temporally precise behavioral window into how real-time AI-MC redistributes cognitive resources. Drawing on Levelt's production model and the psycholinguistics of disfluency, we propose a dual-axis framework: fluency metrics index *when and how much* an AI intervention disrupts production, while lexical metrics reveal *whether* the speaker meaningfully integrated, merely copied, or actively resisted the AI's contribution. We further argue, as a parallel and undertested claim, that AI smoothing of a partner's disfluencies removes the very cues listeners rely on to anchor comprehension, a receptive-side cost that output metrics cannot detect, and we call on the community to study. The framework enables behavioral profiling of human–AI alignment that complements standard NLP evaluation.

1 Introduction

We argue that a new mission for NLP evaluation is to instrument the human production and comprehension pipelines that real-time AI now intervenes in, and that disfluency and lexical entrainment together provide the behavioral instruments that mission requires. The evaluation of spoken dialogue systems has long been anchored to output-side metrics: word error rate (WER) for automatic speech recognition (ASR), BLEU and chrF for machine translation (MT), and task completion for goal-oriented dialogue (Papineni et al., 2002; Post, 2018). These metrics share a common assumption that a better system produces better output, and they have driven substantial progress in speech recognition, spoken language understanding, and end-to-end (e2e) dialogue modeling (Shriberg, 1994; Jurafsky and Martin, 2024). Yet as AI moves from correcting transcripts to live AI-mediated communication (AI-MC) (Hancock et al., 2020), this output-centric view becomes insufficient. Real-time systems that deliver contextual suggestions mid-turn, overlay lexical clarifications on live captions, or generate spoken completions via voice synthesis intervene in the speaker's *production pipeline* while planning, formulating, and articulation are ongoing (Fang et al., 2025; David Ubur, 2025; Xia et al., 2023). The cognitive consequences of this intervention are largely invisible to output metrics: how much it disrupts, where in the pipeline it lands, and whether the speaker meaningfully integrates or merely parrots the AI contribution. An equally invisible cost lands on the listener: when AI smooths the speaker's output, the disfluencies that listeners use to anchor comprehension are stripped from the signal. Together these blind spots motivate a reorientation of NLP evaluation around the speech–cognition interface, not output fidelity alone. This shift is recent and rapid: until 2023, AI in spoken interaction was largely post-hoc (ASR, MT, summarization) and could legitimately be evaluated on output fidelity alone, and live speech-LLM systems are now closing that window faster than the evaluation infrastructure can catch up (Défossez et al., 2024; Zhang et al., 2023; Skantze, 2021). We argue the field needs evaluation instruments commensurate with what these systems do to human speakers, before they are deployed at scale.

Human speech production is a remarkably rapid and largely automatic process. Speakers produce

approximately 2–3 words per second, selecting from a mental lexicon of tens of thousands of entries while simultaneously managing phonological encoding, prosodic planning, and articulatory execution (Levelt, 1989). Given the high cognitive demand, speech disfluencies (e.g., filled pauses, repetitions, restarts) occur in roughly 6–10% of all spoken words in naturalistic conversation (Bortfeld et al., 2001; Fox Tree, 1995). Decades of psycholinguistic research have established that disfluencies are systematic traces of the cognitive processes underlying speech planning, monitoring, and coordination (Goldman-Eisler, 1968; Clark and Fox Tree, 2002; Corley and Stewart, 2008). Crucially, *different disfluency types index distinct production stages*, a property that makes them diagnostic (Shriberg, 1994; Levelt, 1989).

This paper advances a position through a dual-axis framework of *fluency metrics* and *lexical metrics*: when AI intervenes in real-time speech, it creates measurable perturbations in the production pipeline, and the resulting pattern of disfluency (type, rate, temporal position, duration) reveals *where* the perturbation occurred, *how costly* it was, while complementary lexical analysis reveals *whether* the speaker genuinely integrated, superficially copied, or actively resisted the AI's contribution. This claim rests on three well-established psycholinguistic principles: (1) different disfluency types originate at distinct stages of production (Levelt, 1989; Postma, 2000); (2) disfluency rates increase systematically with cognitive load at the relevant stage, up to a moderate-overload regime (Goldman-Eisler, 1968; Oomen and Postma, 2001; Bóna and Bakti, 2020); and (3) disfluencies carry communicative signal for listeners, marking upcoming novelty and processing difficulty (Arnold et al., 2003; Fox Tree, 2001; Corley et al., 2007). The load–disfluency relation we assume is not strictly monotonic: under extreme overload, speakers also *decrease* overt disfluency by going silent or terminating turns prematurely (Oomen and Postma, 2001); the framework therefore applies in the moderate-load regime where intervention costs surface rather than suppress production. Together, fluency and lexical metrics offer NLP researchers a process-level behavioral toolkit that complements output evaluation and opens new empirical questions about human–AI alignment in real-time speech.

2 Disfluency in Speech Theories

Levelt's 1989 model organizes speech production into three incrementally operating stages: the ***conceptualizer*** generates a preverbal communicative intent; the ***formulator*** translates this into a linguistic plan via lemma selection, syntactic encoding, and phonological specification; and the ***articulator*** executes the motor program as overt speech (Indefrey and Levelt, 2004; Levelt et al., 1999). A ***perceptual loop monitor*** runs in parallel, comparing inner speech and auditory feedback against intent; detected mismatches interrupt the ongoing plan and produce overt disfluencies (Roelofs, 2020; Nozari et al., 2011; Hartsuiker and Kolk, 2001). More recent predictive-processing accounts revise this picture by positing that speakers and listeners forward-model upcoming utterances at multiple linguistic levels, with disfluency emerging where prediction mismatches the unfolding signal (Pickering and Garrod, 2013; Hagoort, 2019). We adopt an explicit synthesis: Levelt's serial stages identify *where* in production a perturbation can be localized (conceptualizer, formulator, articulator, monitor), while the predictive account identifies *why* it surfaces, a mismatch between a forward model and the unfolding plan. AI intervention, in this synthesis, injects an external prediction into the forward-modeling loop; the speaker's monitor must reconcile it with the internal plan, and that reconciliation cost surfaces as a disfluency where the mismatch was detected. We treat Levelt's model as the architectural map and prediction-error as the proximal mechanism.

Relation to NLP Research on Disfluency

Disfluency has long been a productive site of NLP research, primarily as a phenomenon to be removed: early work on the Switchboard corpus established taxonomies of interruption points, reparanda, and repairs (Shriberg, 1994), and a subsequent line of neural detection models (from sequence labelers to pretrained transformers and e2e ASR rescorers) treats disfluency as noise to be filtered before downstream parsing or MT (Wang et al., 2020; Jamshid Lou and Johnson, 2017; Rocholl et al., 2021). Parallel lines of work measure AI's effect on language: smart-reply and predictive-text systems narrow lexical diversity and shift register in text-based chat (Arnold et al., 2020; Mieczkowski et al., 2021; Hohenstein et al., 2023; Yakura et al., 2025), while conversational

entrainment in human dialogue provides a baseline for measuring lexical and syntactic convergence (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). Speech is now entering the same regime: e2e models such as Moshi (Défossez et al., 2024) and SpeechGPT (Zhang et al., 2023), and incremental dialogue systems with real-time turn-taking (Skantze, 2021), place AI directly inside the speaker's production loop. Absent from this body of work is the inverse question: how can disfluency patterns serve as a behavioral signal of the *speaker's* cognitive experience and the quality of AI-MC, whether the user is speaking with an LLM directly or with another person through a mediated channel? Our methodological contribution is the reframing of disfluency as a window onto the speaker's real-time engagement with AI-MC.

2.1 Disfluency as Stage-Specific Indicators

The critical insight for the present framework is that different disfluency types predominantly reflect disruptions at different production stages. ***Silent pauses*** at clause and phrase boundaries primarily index conceptual planning difficulty (Goldman-Eisler, 1968; Butterworth, 1980). Their duration correlates with the conceptual complexity and novelty of the upcoming utterance. ***Filled pauses*** ("uh/um") are associated with delays during conceptual planning and lexical retrieval but, importantly, are not merely retrieval artifacts: Clark and Fox Tree (2002) argued that *uh* and *um* are deliberate words that speakers use to signal upcoming delay to listeners, with *um* marking a major delay (often at planning boundaries) and *uh* a minor one. Filled pauses therefore carry a dual signature: a production-side trace of retrieval or planning difficulty, and a comprehension-side signal that listeners can recruit (a duality our §5.2 receptive-side argument depends on). ***Repetitions*** often function as stalling devices that allow the speaker to maintain the conversational floor while formulation catches up with an ongoing conceptual plan (Schegloff et al., 1977; Engelhardt et al., 2010). ***Self-repairs*** indicate that the perceptual loop monitor detected an error (e.g., incorrect words or phonological mismatches) and initiated a halt-and-start sequence (Hartsuiker, 2006; Levelt, 1983). Within-clause covert error detection can occur within a few hundred milliseconds of articulation onset (Hartsuiker and Kolk, 2001; Postma, 2000), but *overt* repairs initiated after the speaker has already adopted an AI-suggested word typically surface later, on the order of a clause or two, as the speaker re-evaluates the adopted material against their original intent.

This stage-specificity means that the *pattern* of disfluency carries diagnostic information about which cognitive processes are taxed. A shift from predominantly fillers toward predominantly repairs, for instance, indicates a change from planning difficulty to an evaluation-and-correction process.

2.2 How AI Mediation Perturbs Production

When a real-time AI system intervenes during spoken interaction, it introduces external content into a production pipeline that is ordinarily driven by internally generated representations. Each type of AI intervention targets a different stage: (1) **Content suggestions** (e.g., topic prompts, factual completions) intervene at *conceptualization*, offering a preformed preverbal message that the speaker did not generate independently. (2) **Lexical and stylistic suggestions** (e.g., word recommendations, register adjustments, tonal reframing) intervene at *formulation*, providing a pre-selected lemma or syntactic frame that competes with the speaker's own lexical access process. (3) **Timing cues** (e.g., turn-taking prompts, connection reminders) intervene at the interface between *planning* and *articulation*, altering when the speaker initiates their turn.

In each case, the AI alternative enters the speaker's forward-modeling loop as an external prediction that competes with the internally generated plan; the perceptual-loop monitor must reconcile the two against communicative intent and decide whether to adopt, revise, or reject the external input. This evaluation step constitutes a dual-task demand: it competes for the same working memory resources that support ongoing production (Oomen and Postma, 2001; Engelhardt et al., 2013). The resulting disruption is therefore a *trace of evaluative processing*, adding cost to both the ***Cognitive Fluency*** (the efficiency of underlying mental processes like lexical retrieval and speech planning) and the ***Utterance Fluency*** (the acoustic smoothness of speech production) (Segalowitz, 2010).

These mechanisms motivate a directional prediction: well-timed AI scaffolds that reduce planning difficulty at the conceptualization stage (e.g., providing a relevant summary before a complex response is needed) can *decrease* filled pauses by alleviating lexical search demands. Conversely, poorly timed or contextually misaligned suggestions can *increase* repairs and repetitions by introducing competing lexical candidates that disrupt

formulation. Our theoretical prediction is *cognitive redistribution*: AI-MC shifts the locus and type of disfluency in ways that reveal which production stages are being supported versus burdened, indicating the degree of thoughtful engagement.

3 A Metric Framework for AI-MC

To evaluate the cognitive and relational impacts of AI-MC, this paper proposes a dual-axis framework combining fluency and lexical metrics. We frame this as a *protocol family* rather than a scalar replacement for WER or BLEU: operationalizations below (window sizes, embedding models, alignment tolerances, baseline definitions) should be task- and corpus-specific. As a blanket convention, every metric is reported as a *deviation from the same speaker's unmediated baseline* (calibration phase or interleaved unmediated trials); raw values are not comparable across speakers because individual disfluency rates vary 5–10× (Bortfeld et al., 2001), allowing the §3.3 typology to distinguish scaffolding from overreliance.

Fluency metrics are incrementally constructed, unobtrusive, and theoretically grounded, allowing researchers to observe cognitive friction when and where it occurs in the production pipeline. But disfluency alone is ambiguous: low disfluency can mean smooth alignment or uncritical overreliance, since users routinely defer to fluent-looking AI suggestions under conversational time pressure (Jakesch et al., 2023; Hohenstein et al., 2023). Lexical metrics are therefore necessary to disentangle cognitive scaffolding from overreliance, distinguishing thoughtful integration from mindless echoing of AI output.

3.1 Fluency Metrics

Fluency metrics capture the *process* of speech production, serving as real-time biomarkers of cognitive load and production difficulty. The following measures form the core behavioral indices:

Disfluency rate by type. Counts and rates (per 100 words or per minute) of filled pauses, silent pauses, repetitions, repairs, and false starts, each indexing different production stages.

Disfluency duration. Temporal extent of each event. Longer filled pauses index deeper lexical search; longer editing phases in self-repairs index more extensive reprocessing (Hartsuiker, 2006).

Positional distribution. Where disfluencies cluster relative to AI intervention timestamps. A spike in filled pauses immediately following an AI suggestion indicates disruption of conceptual planning or lexical retrieval; a later spike in overt repairs (within the same or following clause as the adopted material) indicates that the speaker initially incorporated the suggestion but, on reevaluating it against the original intent, halted and corrected. Pre-uptake disfluency and post-uptake repair jointly operationalize the *Evaluation* axis of §3.3, differing only in *when* the evaluative cost surfaced. The alignment window is task- and stimulus-dependent.

Disfluency trajectory. Evolution of disfluency across a conversation, capturing adaptation effects of whether speakers habituate to AI mediation (declining disfluency) or become increasingly burdened by it (escalating disfluency)

3.2 Lexical Metrics

Fluency metrics reveal that a disruption occurred but not its content-level cause: a speaker might pause and repair because they are resisting an AI suggestion or because they are effortfully incorporating it. We frame the content-level signal as *speaker-to-AI lexical entrainment*, extending the dialogue-entrainment tradition (Nenkova et al., 2008; Niederhoffer and Pennebaker, 2002; Reitter, 2014) to the asymmetric case in which one "interlocutor" is an AI system whose prior output constrains the human's lexical choices.

Moving-average type-token ratio (MATTR). A window-corrected measure of lexical diversity (Covington and McFall, 2010); preferred over raw TTR, which is confounded with utterance length. A declining MATTR under AI-MC suggests the speaker is converging on AI-promoted vocabulary, reducing expressiveness (Hohenstein et al., 2023).

Speaker-to-AI lexical entrainment. How much the speaker's utterance overlaps with the AI's prior suggestion, operationalized via sentence-embedding cosine similarity (e.g., SBERT (Reimers and Gurevych, 2019)), BERTScore (Zhang et al., 2020), or n-gram overlap. Entrainment operationalizes the *Adoption* axis of §3.3; combined with *Evaluation*, high entrainment without evaluation evidence indicates *Mindless Echoing* (overreliance), whereas with evaluation evidence it indicates *Endorsed Adoption* (*effortless scaffolding* when pre-uptake, *effortful compliance*

when post-uptake). Decomposing entrainment into semantic and surface (n-gram) components is a further diagnostic: verbatim adoption is a stronger overreliance signature.

Connotation and register shift. Drift of the speaker's affective, formal, or stylistic profile toward the AI's prior output, operationalized via formality classifiers (Pavlick and Tetreault, 2016), function-word style matching (Niederhoffer and Pennebaker, 2002), and embedding-based affective shifts. Register drift is especially diagnostic when AI reshapes tone for a detected context: a sustained pre-AI baseline register suggests AI is acting as translator, whereas progressively reshaped register persisting into unmediated turns suggests AI is rewriting the speaker's voice.

3.3 Synergy of Fluency and Lexical Metrics

The two families together identify distinct *latent* cognitive states of human–AI alignment. Lexical entrainment estimates the *Adoption* axis (did the speaker take up the AI's contribution?), but cannot tell whether that adoption was evaluated or automatic. The positional distribution of disfluency relative to AI events (§3.1) estimates the *Evaluation* axis (did the speaker show evidence of engaging with the contribution?), but cannot tell what was done with the AI's contribution. Crossing the two latent axes yields a behavioral typology (Table 1) in which each cell corresponds to a single cognitive state, with disfluency and entrainment functioning as the (noisy) measurements that estimate the axes.

To illustrate, Table 2 reports four excerpts from a pilot recording in which speakers carried out a collaborative planning task with on-screen AI-generated lexical and sentence suggestions; disfluency events were manually coded against AI-intervention timestamps. Each excerpt instantiates one cell of Table 1; together they make concrete how the *Adoption* signal (lexical uptake of the AI's contribution) and the *Evaluation* signal (presence and timing of disfluency relative to the AI event) jointly identify the underlying cognitive state.

Reading the excerpts diagnostically: in *Mindless Echoing*, the brief retrieval-onset filler is too short to constitute pre-uptake evaluation and the speaker reproduces the AI's noun phrase verbatim, a classic overreliance signature. *Bypass* shows a clause-onset filler whose temporal alignment is to clause planning, not to the AI event, and the speaker proceeds with their own lemma; the AI is invisible to the production trace. *Endorsed Adoption* (effortful-compliance subtype) shows both pre-uptake (*um*) and post-uptake (the phonological *pack-/pack-aged/packable* repair) evaluation cost, accompanying content-level uptake of the AI's term "packable." *Active Resistance* shows pre-uptake evaluation (silent pause + filler) followed by a substitution that preserves the AI's syntactic frame but overrides its lexical choice, indicating an evaluation that resolved into rejection at the formulator stage.

	No evaluation evidence	**Evaluation evidence present**
Adoption	*Mindless Echoing*: speaker takes up the AI's wording with no pre-uptake disfluency, short uptake latency, and (over the session) narrowing MATTR; the overreliance signature.	*Endorsed Adoption*: speaker takes up the AI's contribution after observable evaluation. Two subtypes by where the evaluative cost surfaces: *effortless scaffolding* (pre-uptake disfluency only) and *effortful compliance* (post-uptake repair).
No adoption	*Bypass*: speaker proceeds with own content; AI is effectively ignored, no measurable cost.	*Active Resistance*: speaker evaluates the AI's contribution and rejects or substitutes, paying dual-task cost.

Table 1: Behavioral typology defined by two *latent* axes: *Adoption* (did the speaker take up the AI's contribution?), operationalized by baseline-relative lexical entrainment (§3.2); and *Evaluation* (did the speaker show evidence of engaging with the AI's contribution?), operationalized by pre-uptake disfluency or post-uptake repair time-locked to the AI event (§3.1). Each cell corresponds to one latent cognitive state. *Endorsed Adoption* admits two surface signatures, depending on whether the evaluative cost surfaces before adoption (*effortless scaffolding*) or as a post-uptake repair (*effortful compliance*).

Because each cell corresponds to a distinct mode of human engagement with AI's contribution (echoed, endorsed, ignored, or rejected), speaker's behavioral profile also provides an *indirect, diagnostic* signal for AI-system evaluation: not a scalar replacement for WER or BLEU, but a way to localize *which* mode of AI behavior produced *which* mode of human cognitive consequence (e.g., a system whose suggestions are mostly echoed without evaluation indicates either excellent intent-matching or excessive user deference; a system whose suggestions are mostly rejected indicates a misalignment that output metrics will miss). This

Typology cell	Speaker utterance (verbatim)	Disfluency profile (production stage indexed)	Speaker-to-AI lexical entrainment
Mindless Echoing	"*um* like rolling suitcase"	Brief filled pause at retrieval onset (weak evaluation evidence: under stricter coding, routine retrieval-onset fillers may be excluded).	Verbatim uptake of AI-suggested phrase; short uptake latency.
Bypass	"*um* I think we can include some basic items"	Filled pause at clause onset (planning delay; not time-locked to the AI event).	No uptake; speaker produces own content.
Endorsed Adoption (effortful-compliance)	"*um* it's a pack-. . . packaged. . . packable backpack"	Pre-uptake filled pause + post-uptake truncation + self-repair (retrieval delay followed by a phonological-neighbor error and recovery); both pre- and post-uptake evaluation evidence present.	Content-level uptake of the AI's term ("packable") with surface-level phonological misalignment.
Active Resistance	"it's. . . *um* but maybe instead of a wool hat use a beanie"	Silent pause + filled pause + revision (pre-uptake evaluation that resolves into rejection at the formulator stage).	Syntactic-frame uptake with lexical substitution (wool hat → beanie); rejects the AI's lexical choice.

Table 2: Four anonymized excerpts illustrating the four cells of Table 1, ordered to match the typology's row-major layout (Adoption row first). The disfluency-profile column links each event to the production stage it indexes and to whether it constitutes evaluation evidence (pre-uptake, post-uptake, or neither); the entrainment column captures the parallel lexical signal. Excerpts are drawn from a pilot recording of speakers responding to on-screen AI-generated lexical and sentence suggestions during a collaborative planning task. The study was approved by the authors' institutional review board (number withheld for anonymous review); content-words have been paraphrased to a different topical domain than the original recording to preserve study anonymity. Disfluency type and position are preserved exactly; events were independently coded by two trained annotators against AI-intervention timestamps, with disagreements resolved by discussion. Italicized tokens mark filled pauses; ". . . " marks silent pauses.

complements the output-side evaluation.

Chance-baseline for lexical overlap. A bias problem holds on the *Adoption* axis: a speaker who would have produced wording similar to the AI's suggestion regardless of mediation will appear "adopting" by chance, biasing them out of *Bypass* and into one of the adoption cells (*Mindless Echoing* or *Endorsed Adoption*). A chance baseline is therefore required: we recommend computing entrainment between the speaker's utterance and *counterfactual* AI suggestions sampled from the same generator on unrelated trials (or, when unavailable, against a held-out distribution of plausible alternatives), and reporting observed entrainment as a deviation from this chance distribution.

Operationalizing the axes. The *Adoption* axis is continuous (entrainment as a baseline-relative deviation, with the chance-baseline correction above). The *Evaluation* axis is binary at the event level (pre-uptake disfluency or post-uptake repair present vs. absent within a task-specific alignment window) and continuous in aggregate. For diagnostic use we recommend either (i) treating the axes as continuous predictors and the four cells of Table 1 as a quadrant labeling of the scatter, or (ii) cutting Adoption at the speaker's within-condition mean $\pm$ 1 SD of the chance-corrected baseline and treating Evaluation as the event-level binary. Without baseline normalization on Adoption and time-locked event coding on Evaluation, the typology is hypothesis-generating; with both, it becomes diagnostic.

3.4 Anticipated Objections

Two objections deserve direct response. (1) *Surveys plus WER suffice.* Surveys are reconstructive and cannot localize cost to specific AI events at the sub-second timescale; WER measures transcript accuracy, not the speaker's pipeline. (2) *Disfluency may index AI-unrelated load.* Temporal alignment to AI-intervention timestamps (§3.1) and the lexical axis jointly constrain attribution: spikes time-locked to AI events and accompanied by lexical adoption are diagnostic; neither signal alone suffices. Pipeline-dependence and speaker-variability concerns are addressed in Limitations.

4 Application Across AI-MC Contexts

4.1 Proactive Content Scaffolding

Systems that deliver contextual nudges, idea suggestions, or factual completions during live conversation (Fang et al., 2025; Xia et al., 2023) (and, in the text-input analogue, smart-reply and smart-compose systems whose convergence effects on language are already documented (Arnold et al., 2020; Mieczkowski et al., 2021)) intervene at the *conceptualization* stage, offering a pre-formed

preverbal message the speaker did not independently generate. When such a scaffold aligns with the speaker's intent, it should reduce planning-stage disfluencies (fewer clause-boundary silent pauses, fewer *um*s). When it conflicts, the speaker must evaluate and reject the alternative, producing a detectable spike in filled pauses or false starts immediately following the AI trigger. Over longer timescales, disfluency *trajectories* can reveal whether speakers habituate to proactive mediation (declining disfluency across a session) or experience escalating burden, a signal that output metrics cannot provide.

4.2 Lexical and Stylistic Adjustment

Systems that recommend word replacements, register adjustments, or tonal reframings (David Ubur, 2025) intervene at the *formulation* stage, offering a pre-selected lemma or syntactic frame that competes with the speaker's own lexical access. The lexical axis is especially diagnostic here. Pickering and Garrod's interactive alignment model (Pickering and Garrod, 2004) and its predictive-processing successor (Pickering and Garrod, 2013) predict that successful dialogue partners converge lexically through automatic priming and shared forward models; computational work on dialogue entrainment provides direct NLP-side methodology for measuring this convergence (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). Under AI-MC, a critical question is whether speakers align primarily with their human interlocutor or AI: tracking lexical convergence between the speaker's vocabulary and (a) the human partner versus (b) the AI system across turns operationalizes this. Power asymmetries are also legible: sociolinguistic research demonstrates that lower-power speakers accommodate more to higher-power interlocutors (Giles and Ogay, 2013). If AI provides formalized suggestions and lower-power speakers adopt them at higher rates (higher entrainment, lower MATTR), this asymmetry documents how AI can entrench relational hierarchies by providing a "safe" register adopted primarily by those who feel least free to express themselves independently.

4.3 Self-Presentation and Emotional Attunement

AI tools that refine tone, suggest empathic phrasing, or rewrite messages for politeness intervene at formulation in service of the speaker's self-presentation goals (Goffman, 1959; Walther, 1996). MATTR provides a direct index: a speaker drawing freely on their own vocabulary produces a higher value, whereas one recycling AI-suggested phrasing produces a lower value. A within-speaker comparison of MATTR in AI-MC versus unmediated conditions quantifies the degree to which AI narrows expressive repertoire. The receptive-side consequence of fluent AI-delivered empathy (the smoothness of delivery undermining perceived genuineness) is taken up as part of the §5.2 argument.

5 Methodological Extensions

5.1 Dyadic vs. Multi-Party Settings

In dyadic conversation, the cognitive dynamics of AI-MC are tractable: entrainment analysis can determine whether the two speakers' disfluency patterns, speech rates, and lexical choices converge over time, and whether AI facilitates or disrupts this convergence (Pickering and Garrod, 2004). In multi-party settings, baseline cognitive load is substantially higher: speakers must track multiple participants' contributions, manage complex floor-holding dynamics, and coordinate overlapping turns (Clark, 1996). AI interventions in this context (real-time meeting summarizers, multilingual translation overlays, group facilitation agents) compound an already taxed cognitive system. The prediction is that the same AI features that reduce disfluency in dyadic contexts may exacerbate it in multi-party ones, because the additional channel of AI input pushes working memory demands beyond the threshold at which incremental production can proceed smoothly. Lexical metrics can additionally track distributional effects: whether AI homogenizes the group's vocabulary (all speakers converging on AI-promoted terms) or differentially affects certain participants, revealing asymmetries in cognitive burden and communicative autonomy.

5.2 Productive vs. Receptive Roles

The metrics proposed above primarily target the productive speaker. However, AI-MC also affects receptive processing, and this is where we make our strongest *call for new empirical work*. A listener whose partner's speech has been "smoothed" by AI loses access to the processing cues that disfluencies normally provide. Psycholinguistic evidence shows that filled pauses orient listener attention to upcoming novel or complex content (Arnold et al., 2003) and facilitate lexical integration (Fox Tree, 2001; Corley et al., 2007), and that listeners ac-

tively compensate for naturally occurring disfluencies in their partner's speech (Brennan and Schober, 2001). Smoothing of partner disfluencies can enter the receptive signal through several distinct channels in current systems: ASR-based live captions and meeting transcripts that drop fillers by default (Wang et al., 2020; Rocholl et al., 2021); speech-to-speech translation pipelines that resynthesize a sanitized version of the source; and emerging e2e speech–language models whose synthesis behavior at fillers is itself a design choice (Défossez et al., 2024; Zhang et al., 2023; Skantze, 2021).

Two adjacent findings support this prediction without yet testing it directly. First, listeners exploit disfluencies not only as signals of speaker effort but as cues that allocate attention to upcoming novel or complex content (Arnold et al., 2003; Fox Tree, 2001) and they adapt their compensatory strategies to a partner's baseline disfluency rate over an interaction (Brennan and Schober, 2001; Corley et al., 2007). Second, AI smoothing is a structurally selective degradation: it preserves the surface signal but strips precisely the cues listeners are documented to recruit, so the comprehension-cost predictions established for other forms of degraded partner input should transfer. To our knowledge, no published study directly measures listener comprehension under naturally disfluent vs. AI-smoothed partner speech in real-time dialogue. We argue that this is a critical gap: the prediction is that smoothing should produce increased comprehension-related disfluency in the listener's subsequent turn (more clarification requests, echoed repetitions, longer planning pauses) and degraded recall of novel material, both of which are operationalizable with the same fluency and lexical metrics introduced above applied to the listener's production. The framework therefore extends a falsifiable prediction to a class of systems whose receptive-side effects have so far been evaluated only with output-fidelity metrics.

5.3 Multimodal Triangulation

Disfluency identifies *that* a cognitive shift occurred but not always *why*; complementary streams strengthen causal interpretation.

Gaze and interaction logs. Timestamps of gaze shifts toward an AI-displayed prompt, combined with logs of which AI intervention was active, establish whether a disfluency event coincides with AI processing, enabling event-related analyses analogous to ERP paradigms.

Physiological measures. EDA and HRV validate that cognitive disruptions correspond to autonomic arousal, distinguishing effortful deliberation (moderate arousal) from cognitive overload (high, stress-associated arousal) (Taptiklis et al., 2023).

Acoustic features beyond disfluency. Speech rate, pitch range, and pause duration complement categorical disfluency counts and are extractable from the same ASR pipeline.

Triangulation is especially important for distinguishing *beneficial* disfluency (genuine deliberation, signaling engagement) from *detrimental* disfluency (overload, signaling that AI-MC has exceeded the available cognitive budget).

6 Design Implications and Conclusion

On the evaluation side, disfluency and lexical-entrainment metrics are unobtrusive, automatable via existing ASR and disfluency-detection tools (Shriberg, 1994; Wang et al., 2020; Jamshid Lou and Johnson, 2017; Rocholl et al., 2021; Taptiklis et al., 2023), and aligned with the millisecond timescale at which cognitive costs accrue. They are natural complements to WER, task completion, and user satisfaction in any AI-MC evaluation protocol.

On the design side, if fluency and lexical metrics can be monitored in real time, the AI system itself can become *adaptive*. A disfluency-aware system could reduce intervention intrusiveness when disfluency rates spike (indicating overload) and increase support when production is smooth but lexically impoverished (indicating over-reliance without deliberation). Such a system would close a feedback loop that current architectures leave open: using the behavioral traces of cognitive mediation to calibrate the mediation itself.

More broadly, this paper argues for a reorientation in how NLP evaluates AI-MC at the speech–cognition interface. Output benchmarks were designed for a world where AI processed text after the fact; as AI moves into the live production pipeline, the field needs instruments commensurate with that pipeline. Disfluency and lexical entrainment offer a theoretically grounded, cognitively interpretable starting point, and one concrete answer to what the missions of NLP research should be when strong general-purpose models are no longer the bottleneck: connecting model behavior to its cognitive and communicative consequences for humans on both sides of the exchange.

Limitations

This paper is a position contribution; its predictions await empirical validation at scale. Several specific limitations follow from this status.

Pipeline dependence. Extracting fluency metrics requires an ASR and disfluency-detection pipeline. Modern e2e speech models (Whisper, SeamlessM4T) and most commercial ASR APIs are trained to suppress fillers and reparanda by default. Applying the framework therefore requires either disfluency-preserving variants (Wang et al., 2020; Jamshid Lou and Johnson, 2017; Rocholl et al., 2021), forced alignment over raw audio, or human annotation; each introduces measurement error and cost that future implementation work must quantify.

Speaker variability. Disfluency rates vary 5–10× across individuals as a function of age, L1 status, working-memory capacity, and clinical conditions (Bortfeld et al., 2001; Engelhardt et al., 2010). The framework is intended for within-speaker contrasts (e.g., mediated vs. unmediated conditions for the same speaker, or trajectories within a session) and is not safe for cross-speaker comparisons without explicit baseline normalization.

Typology under-identification. The 2×2 behavioral typology in §3.3 is built on two *latent* axes (Adoption, Evaluation), each estimated from noisy surface measurements (lexical entrainment and time-locked disfluency, respectively). Measurement error on either axis can still produce cell mis-assignment: the chance-baseline (§3.3) addresses bias on the Adoption axis, and multi-modal triangulation (§4.3) addresses residual ambiguity on the Evaluation axis (e.g., distinguishing a routine retrieval-onset filler from genuine pre-uptake evaluation). The framework is thus a hypothesis-generation device rather than a stand-alone diagnostic.

Theoretical anchoring. The stage-to-disfluency mapping draws on Levelt's serial-stage architecture. Alternative predictive-processing accounts of dialogue (Pickering and Garrod, 2013; Hagoort, 2019) may revise which disfluency types index which processes, and the mapping should be revisited as those accounts mature.

Operationalization is deliberately underspecified. Lemmatizer choice, embedding model, n-gram window, MATTR window size, and AI-event window all depend on language, register, and corpus. We do not prescribe specific implementations because the tooling is moving rapidly; this also means that two studies applying "the framework" may not be directly comparable without explicit operationalization details.

Cross-linguistic scope. Our taxonomy and predictions are derived from primarily English-language psycholinguistic literature. Filled-pause inventories, repair conventions, and clause-boundary cues vary substantially across languages. The framework requires per language (or at least per language family) replication before being applied cross-linguistically.

Feedback-loop contamination. A disfluency-aware adaptive system (§5) may train users to suppress disfluencies over time, eroding the very signal the framework relies on. Periodic re-calibration of the baseline, withheld-intervention control conditions, and longitudinal evaluation are required to detect such drift.

Ethical Considerations

The framework extracts behavioral indices of cognitive load from live spoken interaction, which carries nontrivial ethical risk we want to make explicit.

Speaker-population biases. Disfluency baselines vary systematically across speaker populations: stuttering, aphasia, and Parkinson's-related dysarthria produce disfluency unrelated to AI-MC; L2 speakers produce 2–4× more disfluencies than L1 speakers (Segalowitz, 2010). Naive cross-population application will conflate speaker neurotype, L1 status, and AI system quality, and may penalize tools that disproportionately benefit these populations. Within-speaker designs, stratified analysis, and explicit baseline normalization are prerequisites; cross-population AI-system rankings without these corrections should be avoided.

Surveillance dual-use. Real-time extraction of disfluency and cognitive-load signals from speech is the same affordance commercial vendors of workplace monitoring deploy. We propose these metrics for evaluating AI-MC in controlled research settings, not as a deployable monitoring product. Any field deployment would require informed consent, bystander protection, data-minimization protocols, and explicit prohibition on inference about individual performance.

Privacy and IRB. Applying the framework requires recording and processing speech, which is among the most identifying forms of personal data. Although the extracted features (disfluency counts,

MATTR, embedding similarity) are not directly invertible to content, the audio used to derive them is. Institutional review, secure storage, and on-device feature extraction (where feasible) are appropriate baselines. The pilot recording from which Table 2 draws was collected under an approved institutional review board protocol (IRB; number withheld for anonymous review); illustrative excerpts in this paper have been paraphrased into a different topical domain to further protect speaker anonymity.

Worked-example data. The four excerpts in Table 2 are drawn from a larger Wizard-of-Oz study (IRB number withheld for anonymous review) in which 38 non-native English speakers participated in a 15-minute collaborative design conversation with a confederate via a Zoom-based interface with on-screen AI scaffolds. Participants were recruited from a departmental research-credit pool and compensated with course credit; informed consent was obtained under the approved IRB protocol, and the AI-agent setup was disclosed in post-session debriefing. Audio was recorded, transcribed, anonymized, and manually reviewed; the four excerpts shown here are paraphrased to prevent re-identification. Full study materials, including the conversation script and scaffold content, are available at the project's OSF repository (URL withheld for anonymity).

Power-asymmetry equity. §4.2 predicts that lower-power speakers will accommodate AI suggestions at higher rates. If true, this is not only a measurement artifact but a deployment concern: AI-MC tools may disproportionately reshape the speech of those who already have least latitude to express themselves independently, entrenching rather than flattening conversational hierarchies. Field evaluations of such tools should report stratified outcomes across speakers' positional power, not pooled averages.

Training-data feedback. If AI-MC speech (smoothed by ASR, resynthesized by speech-LLMs, or paraphrased by translation pipelines) is itself ingested as training data for the next generation of systems, the natural disfluency signal this framework relies on will erode at the population level, not only at the level of individual users habituating to a single tool (cf. "feedback-loop contamination" in Limitations). Datasheets and provenance metadata distinguishing naturally produced from AI-MC speech are a prerequisite for sustaining this evaluation regime.

Use of AI Assistants

We used LLM-based writing assistants for surface-level writing refinement, copy-editing, and LaTeX formatting only. No research ideas, framework components, analyses, citations, or arguments were generated or sourced from AI assistants; all conceptual content and bibliography selection were authored by the human author(s). All AI-assisted edits were reviewed and approved by the author(s).

References

Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 128–138, New York, NY, USA. Association for Computing Machinery.

Judit Bóna and Mária Bakti. 2020. The effect of cognitive load on temporal and disfluency patterns of speech. *Target. International Journal of Translation Studies*, 32(3):482–506.

Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147. PMID: 11575901.

Susan E. Brennan and Michael F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Brian Butterworth. 1980. Evidences from pauses in speech, language production: Speech and talk, vol. i.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Martin Corley, Lucy J. MacGregor, and David I. Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Sunday David Ubur. 2025. Ai-driven interpretation of nonverbal communication in ar-enhanced real-time captions: Effects on cognitive load, comprehension, and user engagement. In *Proceedings of the 2025 Conference on Creativity and Cognition*, C&C '25, page 45–51, New York, NY, USA. Association for Computing Machinery.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue.

Paul E. Engelhardt, Martin Corley, Joel T. Nigg, and Fernanda Ferreira. 2010. The role of inhibition in the production of disfluencies. *Memory & Cognition*, 38(5):617–628.

Paul E. Engelhardt, Joel T. Nigg, and Fernanda Ferreira. 2013. Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta Psychologica*, 144(2):424–432.

Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A wearable proactive ai "inner-voice" for contextual nudging. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738.

Jean E. Fox Tree. 2001. Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, 29(2):320–326.

Howard Giles and Tania Ogay. 2013. Communication accommodation theory. In *Explaining communication*, pages 325–344. Routledge.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books, New York.

Frieda Goldman-Eisler. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, London.

Peter Hagoort. 2019. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

Robert J. Hartsuiker. 2006. Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes*, 21(7-8):856–891.

Robert J. Hartsuiker and Herman H.J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, 42(2):113–157.

Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1):5487.

Peter Indefrey and Willem JM Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1):101–144. Towards a New Functional Anatomy of Language.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11).

Paria Jamshid Lou and Mark Johnson. 2017. Disfluency detection using a noisy channel model and a deep neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 547–553, Vancouver, Canada. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing*, 3rd edition. Draft chapters available at `https://web.stanford.edu/~jurafsky/slp3/`.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38.

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. Ai-mediated communication: Language use and interpersonal effects in a referential communication task. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. pages 169–172.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Nazbanou Nozari, Gary S. Dell, and Myrna F. Schwartz. 2011. Is comprehension necessary for error detection? a conflict-based account of monitoring in speech production. *Cognitive Psychology*, 63(1):1–33.

Claudy C. E. Oomen and Albert Postma. 2001. Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, 30(2):163–184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Albert Postma. 2000. Detection of errors during speech production: a review of speech monitoring models. *Cognition*, 77(2):97–132.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Reitter. 2014. Alignment in web-based dialogue: Studies in big data computational psycholinguistics. *Topics in Cognitive Science*. Earlier formulation: Reitter, Keller & Moore (2011), "A computational cognitive model of syntactic priming," *Cognitive Science* 35.

Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency detection with unlabeled data and small bert models. *Preprint*, arXiv:2104.10769.

Ardi Roelofs. 2020. Self-monitoring in speaking: In defense of a comprehension-based account. *Journal of Cognition*, 3(1):18.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Norman Segalowitz. 2010. *Cognitive bases of second language fluency*. Routledge.

Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. *Doctoral dissertation, University of California at Berkeley*.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Nick Taptiklis, Merina Su, Jennifer H. Barnett, Caroline Skirrow, Jasmin Kroll, and Francesca Cormack. 2023. Prediction of mental effort derived from an automated vocal biomarker using machine learning in a large-scale remote sample. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43.

Shaolei Wang, Wanxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Multi-task self-supervised learning for disfluency detection.

Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. Crosstalk: Intelligent substrates for language-oriented interaction in video-based communication and collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. Empirical evidence of large language model's influence on human spoken communication. *Preprint*, arXiv:2409.01754.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.