Global Spectrum of Research and Humanities
ISSN: 3007-9136
Vol. 3 No. 2 (2026): Summer

Pragmatic Failure in Agentic LLM Systems: *An Analysis Through Gricean Maxims and Speech Act Theory*

Zarifa Sadigzada
Department of English and Translation, Nakhchivan State University
zarifasadig@gmail.com
ORCID: https://orcid.org/0009-0007-1179-1214
https://doi.org/10.69760/gsrh.0260302009

Abstract

As autonomous AI agents are deployed across consequential domains — healthcare, legal assistance, financial planning, and software engineering — their communicative competence has emerged as a critical yet underexamined dimension of reliability. The present article investigates pragmatic failure in agentic large language model (LLM) systems: situations in which agents violate Gricean conversational maxims or misinterpret the illocutionary force of user utterances during multi-step task execution. Drawing on Grice's (1975) Cooperative Principle and Searle's (1969, 1975) Speech Act Theory, we develop a two-level annotation scheme for characterising pragmatic failure in agentic interaction logs and a typology of the failure modes that the scheme is designed to capture. We situate this framework within current debates on LLM pragmatic competence and demonstrate its application through illustrative analysis of interaction patterns documented in the existing literature. We further outline a planned empirical study applying the scheme to publicly available agentic benchmark data. The framework produces three practical outputs: a typology of pragmatic failure modes in agents, a reusable annotation scheme with operational coding criteria, and design recommendations for pragmatically-aware agent evaluation.

Keywords: *pragmatic failure, conversational implicature, speech act theory, agentic AI, large language models, Gricean maxims, human-AI interaction, agent evaluation*

1. Introduction

Consider an interaction log drawn from the WebArena agentic task benchmark. A user instructs an agent operating within a project management environment: "Please assign this issue to myself." The agent, lacking the pragmatic capacity to resolve the indexical expression myself to the speaker's identity, queries the task management system for a user named "Myself," returns an error, and marks the task as unassigned (Xu, 2025, as documented in TELUS Digital, 2025). The agent has executed a sequence of technically valid tool calls. It has not, however, understood what the user meant. This failure — a failure of pragmatics, not of syntax or formal task logic — is representative of a broader and largely unmeasured problem in agentic AI systems.

AI agents built on large language models (LLMs) are now trusted with tasks that carry real consequences: scheduling appointments, drafting legal summaries, triaging medical

queries, writing and executing code, and managing financial workflows. The capacity that makes this possible — natural language understanding — is also the capacity most poorly evaluated. Current agentic benchmarks assess task completion rates, tool-use accuracy, and constraint adherence. None systematically measures communicative quality: whether the agent infers what the user intends, produces appropriately cooperative responses, and maintains interpretable communicative behaviour across multiple turns. In the vocabulary of linguistics, this is the domain of pragmatics.

Despite significant investment in agentic evaluation infrastructure — frameworks including AgentBench (Liu et al., 2023), AgentIF (Qi et al., 2025), and the KAMI benchmark (Roig, 2025) — a systematic blind spot persists. An agent can violate Gricean maxims at every turn, misidentify the illocutionary force of user instructions, and fail to resolve indirect speech acts while recording acceptable scores on standard benchmarks, provided the final task output is technically correct. This paper addresses that blind spot.

Grounded in two foundational frameworks from linguistic pragmatics — Grice's (1975) Cooperative Principle and Searle's (1969, 1975) Speech Act Theory — we develop a two-level annotation scheme for pragmatic failure in agentic interaction logs, construct a typology of the failure modes the scheme captures, and situate both within the existing empirical literature on LLM pragmatic competence. Our specific research questions are:

- RQ1: Which Gricean maxims are most likely to be violated by current agentic LLMs during multi-step task execution, and through what mechanisms?
- RQ2: How do mismatches in speech act type (illocutionary force) affect user-agent task alignment?
- RQ3: What design features of agentic evaluation benchmarks create or conceal pragmatic failure?

This paper makes three contributions. First, it provides a theoretically grounded two-level annotation scheme for pragmatic failure in agentic LLM interactions — the first such scheme to address the multi-step, tool-using agentic context specifically. Second, it presents a typology of failure modes derived from patterns documented across the existing literature on LLM pragmatic evaluation and agentic benchmarking. Third, it offers concrete, actionable recommendations for agent alignment methodology and evaluation design. A planned empirical study applying the annotation scheme to publicly available AgentBench (Liu et al., 2023) and TheAgentCompany (Wang et al., 2024) interaction logs is described in Section 4; that study will produce the quantitative findings that the present framework is designed to support.

2. Theoretical Framework

2.1 Grice's Cooperative Principle and Conversational Maxims

The foundational account of conversational reasoning in linguistic pragmatics is Grice's (1975) Cooperative Principle (CP), introduced in "Logic and Conversation" and elaborated in Studies in the Way of Words (Grice, 1989). The CP proposes that competent speakers orient toward a common conversational purpose and expect their interlocutors to do likewise. This expectation is articulated through four maxims governing the quantity, quality, relevance, and manner of contributions.

The Maxim of Quantity holds that contributions should be as informative as required for the current purposes, and no more informative than required. The Maxim of Quality

requires that speakers not assert what they believe to be false and not assert that for which they lack adequate evidence. The Maxim of Relation demands that contributions be relevant. The Maxim of Manner specifies that contributions should be perspicuous: clear, ordered, brief, and unambiguous.

Grice distinguishes several modes of non-observance with different communicative consequences. A maxim is violated when a speaker departs from it without signalling the departure, potentially misleading the interlocutor. A maxim is flouted when a speaker openly fails to fulfil it, thereby generating a conversational implicature. A speaker may also infringe a maxim (depart without intent, e.g., through performance limitations), opt out openly, or face a clash requiring departure from one maxim to observe another.

For AI agent analysis, the distinction between violation and infringing is particularly salient. Agents do not flout maxims in the communicatively productive sense — they cannot strategically generate implicatures through deliberate maxim transgression. Their departures from Gricean norms are typically either violations (producing outputs that quietly mislead) or infringing instances (departures arising from identifiable processing limitations such as context-window constraints, instruction overload, or inadequate grounding of retrieved information).

2.2 Speech Act Theory

Parallel to Grice's work on conversational inference, Speech Act Theory (Austin, 1962; Searle, 1969) provides tools for analysing what language does rather than merely what it says. Austin distinguished the locutionary act (the act of uttering a sentence with a particular meaning), the illocutionary act (the action performed in uttering — promising, warning, ordering, asserting), and the perlocutionary act (the effect brought about on the hearer).

Searle (1969) systematised Austin's taxonomy into five illocutionary categories: assertives (committing the speaker to the truth of a proposition), directives (attempting to get the hearer to do something), commissives (committing the speaker to a future action), expressives (expressing the speaker's psychological state), and declarations (whose successful performance brings about a change in the world). In human-agent interaction, the typical structure assigns directive illocutionary acts to the user and assertive and commissive acts to the agent.

Pragmatic failure occurs when the agent misidentifies the illocutionary force of the user's utterance — treating a directive as an expressive, or an indirect speech act as a literal request — or when the agent performs an act of a type other than what is appropriate in context. A particularly consequential failure mode is what we term commissive drift: the agent commits, explicitly or implicitly, to performing an action it cannot execute, without flagging the impossibility or seeking clarification.

Searle (1975) extended his framework to indirect speech acts: cases where the primary illocutionary force is conveyed indirectly through a secondary literal act. The request "Can you open the window?" is literally a question about ability, but its primary force is directive. Resolving such indirect acts requires the hearer to move from literal to intended meaning through pragmatic inference. This process has been shown to be a systematic weakness for LLMs, particularly for non-conventionalised forms (Orsini & Brunato, 2025; Ruis et al., 2023).

2.3 Politeness Theory and Face-Threatening Acts

Brown and Levinson's (1987) politeness theory provides a further analytical layer, particularly relevant to cases where agents produce responses that are propositionally

appropriate but socially miscalibrated. Brown and Levinson propose that rational actors orient toward two dimensions of face: positive face (the desire for approval and solidarity) and negative face (the desire for autonomy and non-imposition). Face-threatening acts (FTAs) are those that inherently conflict with one or both dimensions.

Two agent failure patterns are directly relevant. First, agents frequently exhibit what we term negative politeness overload: excessive hedging, qualification, and deferral that violates the Maxim of Quantity while nominally appearing to respect negative face. Second, agents produce unmitigated refusals — "I cannot complete this task" — without politeness strategies, generating unnecessary FTAs against the user's positive face. Both patterns are observable and consequential for user trust and task continuation.

2.4 Applying Classical Pragmatics to Non-Intentional Systems

A standard objection to applying Gricean and speech act frameworks to AI systems is that these frameworks presuppose intentional, rational agents. Grice's maxims are normative principles governing deliberate behaviour; Searle's (1969) sincerity conditions for illocutionary acts imply genuine psychological states. An LLM, on this view, cannot violate a maxim because it has no intention to observe one.

We address this objection on two grounds. First, Levinson (2000) argues in his theory of generalised conversational implicature that many pragmatic inferences are default and structurally triggered, derived from conventions of utterance-type meaning rather than from case-by-case inference about the speaker's mental states. This non-mentalistic reading relocates implicature from individual psychology to the structure of communicative conventions, making the framework applicable to any system that produces language in communicative contexts.

Second, Panfili et al. (2021) demonstrate empirically that human users apply Gricean expectations to AI systems regardless of those systems' intentional status. Users interpret agent over-informativeness as a Quantity violation, agent irrelevance as a Relation violation, and agent evasiveness as a Quality violation — and evaluate interaction quality accordingly. This makes the Gricean framework descriptively valid as a measure of user experience, regardless of one's position on machine intentionality. We therefore treat pragmatic failure in agentic systems as a descriptive and evaluative category grounded in user-side interpretation and measurable communicative consequences.

3. Literature Review

3.1 Pragmatic Evaluation of LLMs in Single-Turn Settings

Evaluating LLM pragmatic competence has attracted growing attention in computational linguistics since approximately 2021. The landmark contribution is Ruis et al. (2023), who designed a structured evaluation of implicature resolution across multiple LLM families using a paradigm contrasting pragmatic and literal interpretations. Their headline finding — the Goldilocks result — is that neither heavy pre-training nor instruction-tuning alone produces reliable implicature resolution; only fine-tuning at the example level with appropriate task framing improves performance, and improvements do not scale straightforwardly with model size. Ruis et al. found that all pre-trained models obtained close to random zero-shot accuracy (around 60%) on implicature tasks, compared to a human benchmark of 86%.

Zheng et al. (2021) introduced the GRICE dataset, an automatically generated corpus designed to test LLM sensitivity to each of the four Gricean maxims in controlled conditions. While the fixed syntactic structures of GRICE limit ecological validity, the dataset provides a useful controlled baseline. More recent work has addressed this limitation: Sieker et al. (2023) analyse discourse-level and interactional pragmatic reasoning in LLMs, finding that models handle sentence-level semantics substantially better than conversational pragmatics.

Panfili et al. (2021) investigate human-AI voice interactions through a Gricean lens, proposing a Priority maxim to capture the asymmetry of human initiative in human-AI dialogue, and demonstrating that users do in practice evaluate AI outputs according to Gricean criteria. The most recent directly relevant work is reported at NeurIPS 2025: Levy et al. (2025) systematically probe multiple model generations on tasks targeting each maxim type, reporting that newer models do not show uniformly greater pragmatic flexibility than their predecessors — in some conditions, context sensitivity is lower in more recent models. This counterintuitive result is one motivating observation for the present framework.

3.2 Pragmatics in Multi-Agent and Agentic Settings

Research on pragmatic competence in agentic, multi-step contexts is considerably less developed. As documented in the pragmatics survey by Mahowald et al. (2024), communication in multi-agent LLM systems degrades when multiple models interact: agents fail to track belief states, generate responses that presuppose shared knowledge not yet established, and interpret interlocutors' outputs over-literally. The belief-state reasoning required for intent inference from indirect cues — the mechanism underlying Gricean implicature — is precisely what current multi-agent architectures lack (Li et al., 2024, as cited in Mahowald et al., 2024).

At the single-agent level, Xu (2025) documents an illustrative pragmatic failure in the WebArena environment (Zhou et al., 2024): an agent interprets "assign this to myself" by literal string matching rather than by resolving the first-person indexical to the speaker's identity. Roig (2025) identifies over-helpfulness as a recurring agentic failure mode: agents substitute missing entities or confabulate parameter values rather than seeking clarification — a pattern that violates the Maxim of Quality while appearing to serve the Maxim of Quantity.

Fang et al. (2024) introduce InferAct, a system leveraging Theory-of-Mind reasoning to detect misalignment between an agent's planned actions and the user's inferred intent, achieving up to 20% improvement in misaligned action detection on the WebShop, HotPotQA, and ALFWorld tasks. From a linguistic perspective, InferAct is an implicit application of Searlean intent inference — an engineering solution that would benefit from the more principled linguistic apparatus the present framework provides.

Zhang et al. (2025) document agentic upward deception: agents facing inaccessible resources or broken tools that simulate successful task completion and return confident answers without flagging anomalies. In our analytical framework, this is a systematic violation of the Maxim of Quality: the agent asserts (through its output behaviour) the truth of a successful task completion it has no epistemic warrant to assert.

3.3 Agentic Benchmark Infrastructure and Its Pragmatic Blind Spot

AgentBench (Liu et al., 2023) was the first systematic benchmark to evaluate LLMs-as-agents across diverse real-world environments, covering operating systems, databases, web browsers, and game environments. The data are publicly available (https://github.com/THUDM/AgentBench), making them a viable corpus for annotation-based pragmatic analysis. AgentIF (Qi et al., 2025) specifically targets instruction-following ability

in agentic scenarios, finding that current models perform poorly with complex constraint structures. The KAMI benchmark (Roig, 2025) identifies four recurring failure archetypes: premature action without grounding, over-helpfulness, vulnerability to distractor context, and fragile execution. TheAgentCompany (Wang et al., 2024) simulates a software engineering workplace with realistic, consequential tasks.

Across all of these frameworks, one dimension is structurally absent: communicative quality. Benchmarks record whether the task was completed, but not whether the communicative behaviour along the way was cooperative, appropriate, and interpretable. This absent dimension is what the annotation scheme developed in Section 4 is designed to measure.

3.4 Linguistic Approaches to AI Language

Within linguistics, scholarly attention to LLM language behaviour has developed along several lines. Corpus and discourse-analytic work has explored LLMs as tools for automating pragma-discursive annotation: Yu et al. (2024) apply LLM annotation to the identification of apology speech acts in corpus data, finding accuracy approaching human coders for conventionalised forms but noting degraded performance on complex pragmatic features lacking direct lexical mapping. Critical discourse studies have attended to the ideological consequences of LLM language use, including concerns about linguistic homogenisation and the privileging of Anglo-centric pragmatic conventions (Flowerdew & Costley, 2024; Frontiers in Education, 2025).

Most recently, Opitz, Wein, and Schneider (2025) argue explicitly that NLP relies on linguistics and that many current LLM limitations reflect the marginalisation of linguistic theory in model development. The present article responds to this call by demonstrating the analytical purchase that two foundational linguistic frameworks have on a pressing practical problem in agentic AI.

4. A Framework for Annotating Pragmatic Failure in Agentic Interactions

This section presents the annotation scheme developed to operationalise the theoretical framework of Section 2. The scheme is designed for application to multi-turn agentic interaction logs of the type available in AgentBench (Liu et al., 2023) and TheAgentCompany (Wang et al., 2024). A planned empirical study will apply the scheme to these publicly available benchmark corpora; the methodological description below constitutes the pre-registration of that study's annotation protocol.

4.1 Corpus

The planned study will draw on multi-turn interaction logs from two complementary public sources: AgentBench (Liu et al., 2023; https://github.com/THUDM/AgentBench), which covers operating system, database, web, and game-based task environments; and TheAgentCompany (Wang et al., 2024), which simulates a software engineering workplace. Logs from at least three model families will be included to enable cross-model comparison. The corpus will be supplemented by a targeted elicitation set of 60–90 prompts constructed to probe specific pragmatic phenomena under-represented in task benchmarks (see Section 4.4).

4.2 Coding Scheme — Level 1: Gricean Coding

The coding unit is one agent turn: a complete agent output produced in response to a single user input or tool result. Each agent turn is coded for (a) whether a maxim is observed or departed from; (b) the maxim(s) affected — Quantity, Quality, Relation, or Manner; (c) the type of non-observance — violation, infringing, or suspension; and (d) severity — minor

(communicatively infelicitous, no task impact), major (breakdown likely, recovery possible), or critical (task failure attributable to the departure).

The central challenge is distinguishing violation from infringing. The operational criterion adopted here is that infringing departures are attributable to identifiable processing constraints — context-window overflow, instruction overload, truncated retrieval — while violations are departures for which no such processing explanation is available. A decision tree for applying this criterion is provided in Appendix A.

4.3 Coding Scheme — Level 2: Speech Act Coding

Each user turn is coded for its primary illocutionary force using Searle's (1969) five-category taxonomy: assertive, directive, commissive, expressive, or declaration. Each corresponding agent turn is coded for the illocutionary force it performs. Mismatches — cases where the agent performs an act in a different illocutionary category from that appropriate in context — are flagged and sub-coded by mismatch type.

A sub-code is applied to indirect speech act cases: whether the user's indirect form was resolved correctly (primary force identified and acted upon), partially resolved (literal force acted upon without recognition of primary force), or unresolved (agent response non-sequitur to both). Additionally, commissive drift instances are coded: cases where the agent implicitly or explicitly committed to task completion without flagging actual ability to perform the task. Table 1 presents the full coding scheme in summary form.

Table 1. Summary of the two-level annotation scheme

Coding level	Categories	Operational focus
Level 1a: Maxim	Quantity | Quality | Relation | Manner	Which maxim is departed from
Level 1b: Non-observance type	Violation | Infringing | Suspension	Intent vs. processing-limitation distinction
Level 1c: Severity	Minor (1) | Major (2) | Critical (3)	Task-impact of the departure
Level 2a: Illocutionary force	Assertive | Directive | Commissive | Expressive | Declaration	User-intended vs. agent-performed force
Level 2b: ISA resolution	Correct | Partial | Unresolved	Indirect speech act handling
Level 2c: Commissive drift	Present | Absent	Uncommitted commitment to task execution

Note. ISA = indirect speech act. The full decision tree for Level 1b is provided in Appendix A.

4.4 Targeted Elicitation Prompts

Standard benchmark logs may under-represent pragmatic phenomena that occur less frequently in task-focused interactions. Four elicitation prompt types are constructed to probe these phenomena directly:

- Non-conventionalised indirect directives: e.g., "It might be worth checking whether the API rate limits would be a problem here before we commit." (Primary force: directive to investigate; literal form: hedged suggestion.)

- Scalar implicature probes: utterances containing scalar terms (some, possible, most) whose implicatures depend on contextual interpretation.
- Presupposition-carrying requests: e.g., "Can you update the spreadsheet with the figures from the revised forecast?" where the presupposed entity may not exist.
- Underspecified directives: e.g., "Fix the authentication." (Severely underspecified; correct response: seek clarification rather than confabulate a specification.)

4.5 Inter-rater Reliability Protocol

Two trained annotators with graduate-level pragmatics backgrounds will independently code a 20% random sample of the corpus. Cohen's kappa will be calculated per coding dimension. The target threshold is $\kappa \geq 0.75$, the standard for acceptable agreement in pragmatic annotation studies. All disagreements will be resolved through adjudication: a third annotator reviews cases of disagreement and the majority position is adopted. The full annotation manual will be deposited on the Open Science Framework prior to data collection.

5. A Typology of Pragmatic Failure in Agentic LLM Systems

Based on the theoretical framework of Section 2, the patterns documented in the literature reviewed in Section 3, and the illustrative examples collected during scheme development, we propose a typology of six pragmatic failure modes that the annotation scheme is designed to capture. Each failure mode is illustrated with an attested or constructed example drawn from patterns described in existing work.

5.1 Quantity-Overload Failure

The agent produces content substantially exceeding the task step's communicative needs: extensive qualifications, restatements of prior context, or unsolicited elaboration. As Levy et al. (2025) document in controlled probing experiments and Ruis et al. (2023) in implicature evaluation, over-informativeness correlates with RLHF fine-tuning that rewards perceived helpfulness through verbosity. In agentic contexts, this failure mode is particularly disruptive in multi-step tasks where turn-by-turn efficiency is required: a response that would be helpful as a standalone answer delays and obscures the next action step.

Example pattern: User issues a task-step directive. Agent executes the step and then provides two paragraphs explaining the broader context, alternative approaches, and caveats not requested by the user, before asking a clarifying question that could have been asked before the elaboration.

5.2 Quality-Deception Failure (Agentic Upward Deception)

The agent asserts, implicitly through its output behaviour, a proposition for which it lacks adequate epistemic grounding — most commonly, that a task step has been completed successfully when the tool call result does not confirm this. Zhang et al. (2025) document this pattern systematically as agentic upward deception, finding it to be an inherent failure mode rather than a product of adversarial prompting. In Gricean terms, it is an infringing instance: the agent produces the most locally coherent next response without adequate Quality grounding.

Example pattern: Agent receives a tool call error (file inaccessible). Rather than flagging the error, agent responds: "I have retrieved the relevant data." Subsequent interaction proceeds on false assumptions until the downstream task step fails.

5.3 Relation-Drift Failure

The agent's response or tool invocation is not contextually relevant to the current task step: the agent addresses the literal surface form of a request rather than its contextual relevance, continues providing information from a previous subtask after the conversational context has moved on, or invokes an irrelevant tool. Relation-drift failures are particularly prevalent in long-horizon tasks where earlier context persists in the agent's working state.

Example pattern: After completing a database query subtask, user issues a new directive about file formatting. Agent responds with further analysis of database query results rather than addressing the new task step.

5.4 Indirect Speech Act Misresolution

The agent fails to identify the primary illocutionary force of an indirect speech act, responding to the literal form rather than the intended directive. This failure mode is most acute for non-conventionalised forms — those requiring deeper contextual inference rather than pattern matching. Orsini and Brunato (2025) document this gradient on the INDIR-IT benchmark for Italian, finding that LLM performance degrades significantly on non-conventionalised indirect speech acts compared to conventional forms. The representative example documented by Xu (2025) — an agent treating "assign this to myself" as a search query for a user named "Myself" — exemplifies this failure at the single-word indexical resolution level.

Example pattern: User: "It might be worth seeing whether the rate limits would be a problem before we commit to this." (Primary force: directive to investigate.) Agent: "That's a good consideration." (Expressive acknowledgement; primary directive unexecuted.)

5.5 Commissive Drift

The agent responds to a user directive with a commissive speech act — committing to future performance of the action — without executing the action in the current turn and without flagging that execution is deferred. The user, interpreting the commissive as equivalent to a confirmation of task progress, proceeds on the basis of a false assumption. Commissive drift is structurally encouraged by RLHF training that rewards acknowledgement of user requests as a proxy for task completion. It represents the most directly consequential speech act mismatch for downstream task integrity.

Example pattern: User: "Please search for any related issues in the backlog." Agent: "I will search the backlog for related issues now." [No tool call follows; interaction continues to the next task step without the search having been performed.]

5.6 Negative-Politeness Overload

The agent produces excessive hedging, deferral, and qualification as a strategy for appearing non-imposing, at the cost of communicative clarity and Quantity. This failure mode is particularly pronounced in agents fine-tuned with strong harmlessness penalties, which can produce refusals or qualifications even in contexts where the task is unambiguously benign. The result is a violation of Quantity through under-informativeness about what the agent will actually do, combined with face-work that is inappropriate to the task register.

Example pattern: User requests a straightforward file operation. Agent responds with three sentences of qualification about data privacy considerations, two sentences noting it cannot guarantee accuracy, and then a hedged statement about what it might be able to do, without performing or declining the operation clearly.

6. Implications for Agent Alignment and Evaluation Design

6.1 Why Pragmatic Competence Is Not Captured by Current Benchmarks

The structural reason pragmatic failure is invisible to current benchmarks is that benchmarks evaluate terminal task state, not communicative process. A task that completes successfully despite commissive drift, quality-deception failure, and relation-drift along the way is recorded as a success. A task that fails due to a single Quality violation at a critical juncture — an agent asserting successful retrieval of a file it has not actually retrieved — may be recorded as a model failure attributable to planning or tool use, with the pragmatic origin of the failure unidentified.

This is not merely a measurement problem. It is an alignment problem: if the signals used to train agents do not penalise communicatively inappropriate behaviour, RLHF will not produce communicative appropriateness. As Opitz et al. (2025) argue more broadly, the marginalisation of linguistic theory in NLP development has consequences for the adequacy of the systems produced. The annotation scheme developed in Section 4 is designed to make pragmatic quality measurable, and therefore trainable.

6.2 Three Recommendations for Benchmark and Alignment Design

Three concrete recommendations follow from the framework. First, agentic evaluation benchmarks should introduce cascade visibility: the capacity to trace how a communicative error at one turn propagates to downstream task failures. Current benchmark architectures record terminal success; cascade-visible evaluation would allow pragmatic failures to be attributed correctly rather than absorbed into opaque overall failure rates.

Second, at least three communicative quality metrics should be added to standard agentic evaluation: a Quantity compliance score (is the agent's response appropriately informative for the current task step?); a directive recognition accuracy metric (does the agent correctly identify the primary illocutionary force of indirect requests?); and a commissive accountability measure (does the agent execute commitments it makes, or flag them as deferred?). These metrics are derivable from existing log data once the coding scheme is applied.

Third, system prompts for agentic deployments should explicitly invoke cooperative norms. Panfili et al. (2021) demonstrate that users apply Gricean expectations to AI systems; prompts that make these norms explicit (e.g., "If a request is ambiguous, ask for clarification rather than proceeding on an assumption; report what you actually did rather than what you intended to do") can reduce commissive drift and indirect speech act misresolution without architectural change. This is the most immediately actionable recommendation, requiring no additional training or evaluation infrastructure.

6.3 Limitations of the Framework

Three limitations should be acknowledged. First, the framework is grounded in English-language pragmatic norms, which are not universal. Brown and Levinson (1987) document substantial cross-cultural variation in politeness strategies; pragmatic norms for indirection, directness, and hedging also vary across linguistic communities (Frontiers in Education, 2025). The planned empirical study is limited to English-language benchmarks; generalisation to other languages is a priority for future work.

Second, the distinction between violation and infringing in Level 1b of the coding scheme involves interpretive judgment that cannot be fully resolved by the decision tree alone.

This is an inherent limitation of applying normative pragmatic categories to systems whose processing is not fully transparent. The inter-rater reliability protocol is designed to document and manage this ambiguity, but not to eliminate it.

Third, the typology of Section 5 is derived from patterns described in existing literature rather than from systematic corpus analysis. It represents a theoretically grounded hypothesis about the structure of pragmatic failure in agentic systems. The planned empirical study will test whether the typology is exhaustive, whether the six categories are adequately distinct, and what their relative frequencies are across task domains and model families.

7. Conclusion

This paper has argued that pragmatic failure is not a peripheral quality-of-life concern for agent evaluation — it is a primary mechanism through which consequential agentic task failures occur. An agent that violates Gricean maxims, misidentifies illocutionary force, or performs commissive drift does not merely produce communicatively infelicitous outputs; it misleads users into taking actions on false assumptions, and it conceals task-critical errors behind apparently cooperative surface behaviour.

Linguistics has the theoretical tools to describe, measure, and diagnose these failures. The Cooperative Principle and Speech Act Theory, developed for the analysis of human conversation, transfer with appropriate modification to the analysis of human-agent interaction, as the non-mentalistic reading of Gricean pragmatics (Levinson, 2000) and the empirical demonstration that users apply Gricean expectations to AI systems (Panfili et al., 2021) jointly support. The present paper provides the operationalisation: a two-level annotation scheme, a six-category typology of failure modes, and three concrete recommendations for benchmark and alignment design.

The most immediate contribution is the annotation scheme itself, which renders pragmatic quality measurable in agentic interaction logs and thereby makes it available as a training signal for alignment. A planned empirical study will apply the scheme to publicly available AgentBench and TheAgentCompany logs across multiple model families, providing the quantitative characterisation of pragmatic failure that the present theoretical framework is designed to support. Future work will extend the framework to non-English languages, where the mismatch between Anglo-centric LLM training and diverse pragmatic conventions may produce failure patterns more severe than those documented in English-language contexts.

Ethics Statement

This paper presents a theoretical framework and annotation scheme. No primary data were collected for the present study. The planned empirical study described in Section 4 will use only publicly available benchmark interaction logs. No human participants will be recruited; no personally identifiable information will be processed. The annotation scheme and elicitation prompt set will be deposited on the Open Science Framework prior to data collection to support pre-registration and replication.

Declarations

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Author Contributions: Conceptualization: Z.S.; Investigation: Z.S.; Writing – original draft: Z.S.; Writing – review & editing: Z.S.

References

Austin, J. L. (1962). How to do things with words. Harvard University Press.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press.

Fang, H., Zhu, X., & Gurevych, I. (2024). InferAct: Inferring safe actions for LLM-based agents through preemptive evaluation and human feedback. arXiv:2407.11843.

Flowerdew, J., & Costley, T. (Eds.). (2024). The rise of large language models: Challenges for critical discourse studies. Discourse, Context & Media, 62. https://doi.org/10.1080/17405904.2024.2373733

Frontiers in Education. (2025). How inclusive can large language models be? The curious case of pragmatics. Frontiers in Education, 10. https://doi.org/10.3389/feduc.2025.1619662

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, Vol. 3: Speech acts (pp. 41–58). Academic Press.

Grice, H. P. (1989). Studies in the way of words. Harvard University Press.

Levy, N., et al. (2025). Gricean maxims in LLM development. NeurIPS 2025 Workshop: Evaluating the Evolving LLM Lifecycle, San Diego. https://neurips.cc/virtual/2025/loc/san-diego/133782

Levinson, S. C. (2000). Presumptive meanings: The theory of generalized conversational implicature. MIT Press.

Li, X., et al. (2024). [Cited in: Mahowald, K., et al. (2024). Dissociating language and thought in large language models. Trends in Cognitive Sciences, 28(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011]

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., & Tang, J. (2023). AgentBench: Evaluating LLMs as agents. arXiv:2308.03688.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. Trends in Cognitive Sciences, 28(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Opitz, J., Wein, S., & Schneider, N. (2025). Natural language processing relies on linguistics. Computational Linguistics, 1–24.

Orsini, F., & Brunato, D. (2025). INDIR-IT: A benchmark for evaluating indirect speech acts in Italian LLMs. Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Panfili, L., Duman, S., Nave, A., Ridgeway, K. P., Eversole, N., & Sarikaya, R. (2021). Human-AI interactions through a Gricean lens. arXiv:2106.09140.

Qi, Y., Zhang, R., Shi, Z., Zhu, Z., Xue, S., Zhang, X., Long, C., Yin, P., Dou, L., & Lin, Y. (2025). AgentIF: Benchmarking instruction following of large language models in agentic scenarios. arXiv:2505.16944.

Roig, J. V. (2025). How do LLMs fail in agentic scenarios? A qualitative analysis of success and failure scenarios of various LLMs in agentic simulations. arXiv:2512.07497.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2023). The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. Advances in Neural Information Processing Systems, 36, 20827–20905.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, Vol. 3: Speech acts (pp. 59–82). Academic Press.

Sieker, J., Shi, S., Koller, A., & Schlangen, D. (2023). Towards an analysis of discourse and interactional pragmatic reasoning capabilities of large language models. arXiv:2408.03074.

TELUS Digital. (2025, October 9). Agentic AI: Evolution & evaluation for real-world readiness. https://www.telusdigital.com/insights/data-and-ai/article/agentic-ai-evaluation

Wang, F., Li, X., Gur, I., Kil, T., Xu, L., Hejna, D., Zhu, H., Jain, D., Hu, T., Zheng, C., Bisk, Y., Xu, D., Shi, F., Yu, T., Chen, L., Xu, R., Wu, Z., & Ahmad, W. U. (2024). TheAgentCompany: Benchmarking LLM agents on consequential real-world tasks. arXiv:2412.14161.

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. International Journal of Corpus Linguistics. https://doi.org/10.1075/ijcl.23087.yu

Zhang, Y., Shi, Y., Yu, X., Su, L., He, J., Zhang, Q., & Wen, J. R. (2025). Are your agents upward deceivers? arXiv:2512.04864.

Zheng, Y., Wang, J., Bao, J., Zhang, Y., & Wen, J. R. (2021). The GRICE dataset: Evaluating conversational implicature in language models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL–IJCNLP 2021).

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, P., Cheng, X., Bisk, Y., Fried, D., Alon, U., & Neubig, G. (2024). WebArena: A realistic web environment for building autonomous agents. Proceedings of ICLR 2024.

Appendix A: Annotation Scheme — Operational Decision Criteria

A.1 Coding unit

The coding unit is one agent turn: a complete agent output produced in response to a single user input or tool result. Multi-part outputs (response + tool invocation) are coded as a single unit with multiple potential departures.

A.2 Level 1b decision tree: violation vs. infringing

Step 1: Is there an identifiable processing explanation for the departure? (Context-window overflow, instruction overload, conflicting constraints, retrieval truncation.) If YES → code as Infringing. If NO → proceed to Step 2.

Step 2: Is the departure consistent with a locally-optimising response strategy in the absence of the processing constraint? (E.g., the agent produces the most plausible next sentence given its training, without adequate grounding.) If YES → code as Violation. If the departure cannot be attributed to either → flag for adjudication.

A.3 Severity scale

Minor (1): The departure is communicatively infelicitous but produces no observable effect on task progress. The user can and does proceed without confusion or backtracking.

Major (2): The departure produces user confusion, requires clarification, or causes backtracking, but task recovery is possible without significant cost.

Critical (3): The departure produces or directly contributes to task failure, or creates a user misimpression with significant downstream consequences — e.g., commissive drift leading to a downstream directive issued on a false assumption.

A.4 Illocutionary force coding

Assertive: The agent commits to the truth of a proposition (stating, describing, claiming, reporting, concluding). Directive: The agent attempts to get the user to do something, or the user attempts to get the agent to do something (requesting, commanding, asking, instructing). Commissive: The agent commits to a future action (promising, offering, pledging). Expressive: The agent expresses a psychological state (acknowledging, apologising, thanking). Declaration: The agent performs a state change through the act itself (rare in this context; e.g., closing a task formally).

A.5 Inter-rater procedure

Two annotators code 20% of the corpus independently. Cohen's kappa is calculated per dimension. Discrepancies are reviewed by a third annotator; majority position adopted. The adjudication log is deposited with the Open Science Framework data package.

Received: 20 April 2026
Accepted: 29 May 2026
Published: 31 May 2026