العراقية
المجلات الاكاديمية العلمية
IRAQI Academic Scientific Journals
Al-Itqan Journal of Humanities
available online at
https://iasj.rdd.edu.iq/journals/journal/view/575
ISSN: 3105-1073 (OnIne)
المجلة معتمدة بالترقيات العلمية حسب كتاب وزارة التعليم العراقية (46404-1410) في 2026/4/27

A Pragmatic Analysis of Failure in AI-Generated Discourse: Misinterpreted Contexts and Communicative Intentions

Israa Sabeeh Abbas, M.A.
israa.s.alfalahi@aliraqia.edu.iq
israa.sabeh@gmail.com
Instructor in Department of English language
College of Education for Women
AL- Iraqia University

Abstract

Large language models generate text that is almost grammatically faultless, yet that generated text often fails to function appropriately within its context. Jenny Thomas identified this misfunction as pragmatic failure, i.e. a lapse or a breakdown in grasping what a speaker means rather than a grammatical defect. This article investigateshow pragmatic failure manifested in discourse produced by AI. The article also interrogates whether the concept still applies when the speaker has no communicative intentions at all. Methodologically, the study is conducted as a conceptual, integrative review. It draws on four areas: Theories of pragmatic failure, Gricean analyses of human-machine talk, work on theory of mind and intent research, and pragmatic account of hallucination. Reported failures are coded using a combined Gricean and Thomasian framwork. Five patterns emered of this. Relevance failures are the biggest source of user dissatisfaction across languages and platforms. Comprehension failures cluster where interpretation must reconstruct unstated context. Sociopragmatic failure is most visible outside English. Hallucination acts like a failure of warrant, not just a factual error. Pragmatic failure in machine talk is doubled because oneside of conversation has no real speaker. The study argues for redefining the concept: seprating the failures of recovery from failures of warrant. It treats the human-machine process as the unit of analysis. The revision builds on a shift that cross-cultural pragmatics had already anticipated.**Keywords:** pragmatic failure, large language models, Gricean maxims, human–AI interaction, hallucination,AI- generated discourse

المستخلص

تنتج النماذج اللغوية الكبيرة نصوصًا شبه خالية من الأخطاء النحوية، غير أن هذه النصوص كثيرًا ما تخفق في مطابقة السياق. وهذا النمط هو ما سمّاه توماس بالإخفاق التداولي، أي القصور في إدراك ما يقصده المتكلم لا الخطأ في القواعد. تسأل هذه الدراسة كيف يتبدّى هذا الإخفاق في الخطاب المولَّد بالذكاء الاصطناعي، وهل يصمد المفهوم عند احتكاكه بمُحاوِر لا يقصد. تعتمد الدراسة تصميم المراجعة التكاملية المفاهيمية، وتجمع دراسات أولية من أربعة حقول: نظرية الإخفاق التداولي، والتحليلات الغرايسية للتخاطب بين الإنسان والآلة، وأبحاث نظرية العقل والقصد، وإعادة التأطير التداولي لظاهرة الهلوسة. رُمِّزت حالات الإخفاق المرصودة وفق شبكة تجمع بين معايير غرايس وتصنيف توماس. برزت خمسة أنماط. تهيمن مخالفات المناسبة على استياء المستخدمين عبر اللغات والمنصّات. وتتجمّع إخفاقات الفهم حيث يتطلب التأويل إعادة بناء سياق غير مُصرَّح به. ويظهر الإخفاق التداولي الاجتماعي بوضوح أكبر خارج الإنجليزية. وتعمل الهلوسة بوصفها إخفاقًا في المسوّغ لا في الحقيقة. والإخفاق التداولي في خطاب الآلة مُزاح إزاحة مضاعفة، لأن أحد قطبي الثنائية التواصلية خالٍ. تدعو الدراسة إلى إعادة تحديد المفهوم، بالفصل بين إخفاقات الاسترجاع وإخفاقات المسوّغ، وبمعاملة العملية المقترنة بين الإنسان والآلة وحدةً للتحليل. وتمثّل هذه المراجعة امتدادًا لإعادة تعريف سبق أن استشرفتها التداولية عبر الثقافية.الكلمات المفتاحية: الإخفاق التداولي؛ النماذج اللغوية الكبيرة؛ معايير غرايس؛ التداولية الاجتماعية؛ القصد التواصلي؛ التفاعل بين الإنسان والذكاء الاصطناعي؛ الهلوسة؛ المناسبة.

1. Introduction

The insight that initiated this study of pragmatic failure has developed unexpected second trajectory. Jenny Thomas (1983) drew a clear line between grammar and use. She confirmed that pragmatic failurehas not stem from grammatical mistakes or errors. For almost twenty years, that line mostly separated two kinds of learners: one who makes grammatical and another who builds flawless sentences that nevertheless fail pragmatically. The latter kind, though rare, proves instructive.Large language models (LLM) flipsThomas's ratio. Today, the flawless sentence that still misses the point is now the ordinary case. Studies of model performancereveal the same split: accurate grammatical utterance, inconsistent context- sensitive language (Mahowald et al., 2024). A system that almost never violates grammar, yet often misinterprets the purpose of an utterance aligning with Thomas'scharacterization of pragmatic failure This correspondence is a direct application of pragmatic theory. Pragmatic failure, on the relevance-theoretic accounts, occurs when a hearer mistakes the force behind an utterance other than the speaker intended (Zhang, 2021). Both sides of Human- AI interaction exhibit this mismatch. A user may misunderstand achatbot and vice versa. They appear equal in this respect, but they are not on the pragmatic scale. Users may infer intentions the model cannot possess while the AI model constructs anillocutionary force a user never made. By Thomson criterion both cases are classified as pragmatic failure. In her account the intention precedes the interpretation. Human- AI dialogues unsettle this arrangement because one side of the exchange may have no intention of recovering.The article addresses three issues:First, what forms or varieties of pragmatic failure that occur in AI generated discourse, and how do they map into established pragmatic mechanisms? Second, where in the exchange does failurearise, basically in comprehension, or in production, or in the joint process between them? Third, does the inherited concept itself need to be revised for communication that lacks intention The contribution, her, is conceptual and theatrical rather than experimental. Instead of gathering new data, the researcher reexaminedfindings from four research traditions. Bringing those studies together exposes patterns which are otherwise hard to recognizeto assess whether the concept requires revision. Section 2 is a review for the four research traditions. Section 3 presents the framework and adopts Grice and Thomas as the analytic apparatus. Section 4 describes the method. Section 5 reports the findings. Section 6 discusses them. Section 7 concludes the limits and the future research directions.

2. Literature Review

2.1 Pragmatic failure

Scholarship on pragmatic failure still relies on the structure proposed by Jenny Thomas (1983). On her account, pragmalinguistic failure occurs when a linguistically acceptable form of expression carries unintended communicative force. In contrast, sociopragmatic failure stems from an incorrect assessment of what is socially acceptable in a particular interaction (Thomas, 1983; Zhang, 2021). This distinction separates problems of linguistic realization and problems of social interpretation. A speaker (or user) may use the wrong expression for a request, or the right expression in the wrong social context. What fails in both is not the linguistic form but pragmatic competence. Relevance theory, later, restated this idea without abandoning its core. On this reading, pragmatic failure is a hearer misunderstand a speaker's meaning. The cause may lie in the speaker's inappropriate utterance or in the hearer's misinterpretation, or in both (Zhang, 2021). Under this formulation, responsibility is distributed between the two sides while retaining the speaker's meaning as the standard. The concept was subsequently extended into intercultural pragmatics Intercultural pragmatics took the idea carried the idea outward. Pragmatic failure denotes people from different cultural backgrounds cannot use or read is the inability of interlocuters from different cultural backgrounds to use and interpret appropriately in social contexts.The consequences ranges from misunderstanding and misinterpretation to conflict Some studies go even further in touch the taxonomy, placing behavioural and non-verbal failure next to the original two (Ren et al., 2016).All of these assumptions share a single presupposition that both interlocuters are shaped by their cultures. The speaker has a meaning and the hearer has a frame. Failure is the gap between them. Cross-cultural pragmatists had already begun to doubt that this picture holds under current conditions. With English as a lingua franca, and with messy online interaction, they argued, the very notion of pragmatic failure may need redefinition for more diverse norms and more variable judgements (Peter, 2019). That call predates the present systems which render it concrete.

2.2 Human-AI communication

The most usable vocabulary for machine breakdown comes from Grice. His Cooperative Principle, with the maxims of quantity, quality, relation, and manner, is now the default grid for sorting conversational-agent errors. The record built on it is large enough to show stable patterns (Krause & Vossen, 2024).Across the evidence, themost consistent pattern is failure of relevance. Panfili et al. (2021) analyzed 700 voice-assistant interactions. Their analysis revealed that off-topic responses were the dominant error and were judged by usersas the least

natural and the most frustrating. An identical pattern appears in a Korean study. Analyzing more than 3,000 adjacency pairs involving the Kakao Mini assistant showed that the violations of the maxim of relation were both the most frequent and the least acceptable to users (Nam et al., 2023). The convergence extends further to multi-system comparisons against human baseline. Human- AI comparisons (Aziz, 2025) and early test turning research Turing-test research reinforce the recuring of the same problem (Sayginn&Çiçekli, 2002).Convevance extend languages, systems and evaluation settings where relevance emerges as the main source of pragmatic failure.Failures of the remaining maxims are more circumscribed. With respect to quantity, users were dissatisfiedmore stronglywith insufficient information than excessive details. Generally, they were tolerant to resent too little information more than too much. They were tolerant of verbose responses from machines than those of human interlocuters (Panfili et al., 2021). Findings on quality presents different pattern. Some systems maintain high levels of factual accuracy whereas broader chatbot research identifies unverified (inaccurate) content that weakens user trust (Panfili et al., 2021; Aziz, 2025). Manner violations are mainly associated with complaints aboutclarity, yet they remain subordinate to relevance (Panfili et al., 2021). Experimental research highlights an additional processing cost due to the factor that those violations increase the process time and reads as less human (Jacquet et al., 2019). These failures matter and literature started to examine them on their own terms.The remaining literature shifts the focus from relevance to the other dimensions of cooperative communication. The model performance assessments are mixed. Some of them portray current models as largely cooperative (Firdaus et al., 2025). Others, like Aziz (2025) using human baselines and real data, identify widespread violations of all fourmaxims (Aziz, 2025). The disagreement is largely methodological. The models handle explicit prompts effectively but struggle once meaning must be inferred rather than stated directly (Firdaus et al., 2025; Jacquet et al., 2019). Where the frameworks themselves are heading is a separate question, taken up next.The grid has also been pushed into new territory. One line reads the conditions for responsible assertion as continuous with Gricean quality, once fallible knowledge is taken seriously. Another asks whether models trained on child-scale data acquire any sensitivity to the maxims at all. Both moves treat the Gricean grid less as a settled inventory than an instrument still being adapted to new kinds of speakers.A common response to these findings is to reconsider the Gricean framework. Researchers argue that maxims of quality, quantity relation and manner may not be enough for human-AI communication. The most comprehensive proposal introduces benevolence and transparency. Gricean version presumes an interlocuter whose intentions can be inferred.This presupposition fails where AI systems obscure their limitations and objectives (Miehling et al., 2024). Studies of pedagogical agents reach a similar point through trust. Other researchers move beyond evaluation and treat the maxims as design principles rather than analytical criteria. Agents developed with these principles communicate more clearly and respond more accurately. They remain more relevant to the user's request. Participatory studies refine these principles into actionable design guidelines(Murukannaiah & Singh, 2025;Kim et al., 2025).Across these studies, pragmatic success emerges less as a matter of rule compliance than as a content-sensitive interpretation.

2.3 AI and context interpretation

The previous section examines how users evaluate AI- generated responses. Recent literature turns to the interpretive process preceding those responses or more preciselystudy how the model understands what the user says to it. That literature matters because many failures of relevance begin before a response is generated. Hence, a response can only be relevant if the model correctly interprets the purpose of the preceding turn. Evidence from theory of mind research reflects this distinction. Some testing reported near human performance on classical mentalizing tasks like false-belief stories (Kosinski, 2023). Systematic comparative studies recognize a consistent limitation. Models perform less reliably when interpretation requires social knowledge that is not explicitly stated with Faux pas(social blunder) detection emerges as a recurring difficulty (Strachan et al., 2024). Recognizing a Faux pas requires a sociopragmatic judgement. The task is to determine whether a well-formed utterance violet a norm breaks a norm, given the participants, their relationships and their shared knowledge. A model that succeeds on explicittasks but fail Faux pas detection appears better at tracking prepositional contentthan at interpreting the social context in which that content is used. The second ability has always been centered That second layer is the one pragmatic failure has always been about. Studies on intent recognition follow similar patterns. Models identify common user intent but decline as intention becomes less frequent and more complicated (Bodonhelyi et al., 2024). Spoken language adds another layer of complexity to the issue of pragmatic failure. Noice introduced duting recognition reduces the accuracy of intent and slot filling. Therefore, the model responds to a distorted utterance rather than the user's original one (He&Garner, 2023). Here, a failure of interpretation begins as a failure of perception.The different starnds of research come together in conversational implicature. Understanding an implicature requires more than decoding words. It requires the hearer to infer what the speaker intended but left

unsaid. Success, accordingly, depends on context rather than . Work on conversational implicature in human-AI interaction studies exactly this (Salman & Matrood, 2025). A broad survey makes structural complaints. Benchmarks under-cover implicature, reference, and real-world use, and richer datasets are needed (Ma et al., 2025). General reviews echo the unease about evidence resting on narrow tasks (Liu et al., 2023). Here, the challenge lies in what remains unsaid.Sociopragmatic failure surfaces mainly outside English. The studies that go there are not a footnote. They are where Thomas's second category becomes visible in machine talk. Korean evaluation documents over-formality and honorific misuse. The honorific grammar is handled. Its calibration to relationship and status is not (Park et al., 2024). Arabic evaluation documents literalism, where local convention licenses non-literal readings, and dialect mixing that ignores the situation (Alshraah et al., 2025). Literalism is pragmalinguistic. Dialect mixing is sociopragmatic. The same study notes how little cross-linguistic comparison exists across power and familiarity settings (Alshraah et al., 2025). The sociopragmatic picture is still incomplete and needs more exploration.A fourth strand reframes hallucination. The dominant surveys treat it as factual error. They define it as fluent, authentic-sounding text that departs from source inputs or from real-world facts, and sort cases as input-conflicting, context-conflicting, or fact-conflicting (Ye et al., 2023; Aditya, 2024; Huang et al., 2023). They frame it as a threat to trust and reliability (Zhang et al., 2023). One line makes pragmatic reading explicit. Through speech-act theory, model outputs are assertions issued without the conditions that would license them: authority, evidence, situational fit. A proposed pragmatic layer would gate when text may count as an assertion (Li, 2025). Adjacent work treats hallucination as distributed cognition, where humans and AI co-produce shared false versions of events (Osler, 2026). Other work recasts it as a by-product reframable as creative flexibility (Hamid, 2024), or turns to the metaphor itself, arguing that the word anthropomorphises models and spreads responsibility thin (Förster & Skop, 2025). Hallucination therefore reaches beyond factual accuracy.

2.4 Research gap

Three gaps recur. The first is conceptual. The theory of pragmatic failure presupposes two intending, situated parties. None of the empirical work confronts what happens to the concept when one party does not intend. The redefinition that cross-cultural pragmatics called for has not been carried out for synthetic interlocutors (Peter, 2019).The second gap is methodological. Most testing runs in English, in text, under controlled conditions. Reviews of human-computer pragmatics, and of psychological uses of models, converge on one absence: standardised, transparent, cross-cultural benchmarks for pragmatic and social meaning (Quan & Chen, 2024; Abdurahman et al., 2024; Ma et al., 2025). The available evaluation remains methodologically uneven.The third gap concerns real-time dynamics. Pragmatic drift over long conversations, and behaviour under genuine ambiguity, stay under-characterised beyond small tests (Strachan et al., 2024; Liu et al., 2023). This study addresses the first gap directly while treating the other two as direction for future inquiry.

3. Theoretical Framework

The analysis uses two frameworks together. Grice supplies a vocabulary for the texture of cooperative and uncooperative talk. Thomas supplies the categories and the participant structure within which failure is defined. Each is summarised here as an instrument, with attention to what each assumes when it enters the machine case.

3.1 Grice's Cooperative Principle

Grice argued that participants behave as though committed to the purpose of the exchange. He confirms that commitment runs through four maxims. Quantityrequires speakers to provide enough information, but no more than necessary and no more. Quality asks to avoid unsupported claims. Relation requires relevance. Manner needs clarity and order. The maxims are not rules of politeness or etiquette. They describe what people normally expect in talks. When those expectations are not met, hearers begin to look for meaning beyond the words themselves. Hence, the same framework is now used to classify failures in conversational AI. The record shows them doing so reliably, with relevance the most salient dimension (Panfili et al., 2021; Nam et al., 2023; Krause & Vossen, 2024). The maxim identifies the point of breakdown before they explain its source.The inference a hearer draws from an under-informative turn depends on crediting the speaker with a reason for holding back. That inference rests on an assumption. One assumption breaks down in the machine case.Flouting presupposes a cooperative speaker. A model that under-informs is not flouting in that sense. There is no intending speaker from whom the implicature can be recovered. The maxims stay accurate about what users feel. They become unreliable about what is happening on the other side. The proposed maxims of benevolence and transparency partly admit this. They convert an assumption about the partner's mind into a requirement on the system's behaviour (Miehling et al., 2024).

3.2 Thomas's model of pragmatic failure

Thomas contributes the categories and the dyadic structure. The central distinction separates pragmalinguistic from sociopragmatic failure, that is, the mishandling of form-to-force mapping from the mishandling of social appropriateness (Thomas, 1983; Zhang, 2021). The central mechanism is the gap between what a speaker means and what a hearer takes the force to be (Zhang, 2021). The model is useful here for a plain reason. It names the two failure types that the cross-linguistic evidence instantiates almost exactly. Literalism is pragmalinguistic. Honorific miscalibration and dialect mismatch are sociopragmatic (Alshraah et al., 2025; Park et al., 2024)The model also invistgates the assumption explored in this study. It is built around two intending participants. That symmetry disappears in conversations with AI. The user can misunderstand the model, yet no communicative intention exists behind output. The model's interpretation cannot be treated as a human interlocutor's misunderstanding. The framework still captures the forms of failure.

3.3 A combined analytic grid

Rather than operating separately the frameworks are fused into a two-axisanalytical scheme. The Gricean dimension categories failure by the maxim they disturb. It describes and characterizes the immediate experience of the breakdown. The second axis (Thomas's model) classifies failure by pragmatic category and by their position within the speaker- hearer relationship.Applying both models reveals two aspects of each failure at the same time. It identifies what cooperative expectation is violated. It also locates where the disruption of the source failure and the presence or absence of an intending agent. This integrated framework provides the basis for this study.

4. Methodology

This study uses a conceptual, integrative-review design. The design suits a project whose aim is to consolidate evidence from different research traditions and to revise a theoretical construct, rather than to estimate an effect or test a hypothesis against new data. The construct under revision is pragmatic failure. The traditions are the four literatures reviewed in Section 2. No primary data were generated. The synthesis is interpretive, not statistical. Claims about frequency are reported as the cited studies report them. They are not aggregated quantitatively.The corpus comprises peer-reviewed articles, conference papers, and preprints from four areas. The first is the theory of pragmatic failure and its intercultural extensions. The second is Gricean analysis of human-machine and chatbot conversation. The third is theory-of-mind, intent-recognition, and implicature research on language models. The fourth is pragmatic and speech-act reframing of hallucination, together with the survey literature that supplies its taxonomies. Sources were chosen for direct relevance to the meeting point of pragmatic theory and machine-generated discourse. Priority went to studies that report observed failures, propose categories, or argue for conceptual revision. Foundational theory was included for the constructs it supplies, even where it predates the technology, because the framework is the object of revision, not the technology.Each source was read analytically. Its reported failures were coded against the grid from Section 3.3. Coding ran in two passes. The first pass assigned each failure to a Gricean maxim. Where a source framed a failure in those terms, that framing was kept. Where it did not, the failure was assigned to the maxim its description most closely instantiated. Split cases were recorded as splits, not forced into one cell. The divided evidence on quality, for instance, was retained as a division rather than resolved (Panfili et al., 2021; Aziz, 2025). The second pass assigned each failure to a Thomasian type, pragmalinguistic or sociopragmatic. It then located the failure in the dyad. A failure of recovery is the model misreading an intended human meaning. A failure of warrant is the model producing a force beyond anything it is positioned to mean. The five findings in Section 5 are the themes that recurred under this coding. They are presented as patterns in the evidence. Their interpretation is held back for Section 6.Two limits should be stated at the start. The synthesis is shaped by the coverage biases of the available literature. That literature is mostly English-language, mostly text-based, and concentrated on a few widely studied systems. The findings inherit those biases and are read in their light. The coding is interpretive at the points where sources did not use the categories applied. Those points are flagged in the findings where they bear on the argument.

5. Findings

5.1 Relevance is the primary site of breakdown

Across platforms, languages, and designs, relevance violations are the most frequent failure and the most damaging to user satisfaction. As Panfili et. Al. state "Relevance violations were the most frequent of all violations, and they appear to be the most frustrating" (2021, p. 292). This pattern is reinforced by their broader observation that "Human take conversational priority over AI" (p.297). The pattern extends over English voice-assistant data, Korean language interactions, multisystem evaluation, and early Turing-test research (Panfili et al., 2021; Nam et al., 2023; Aziz, 2025; Saygin&Çiçekli, 2002). The evidence leaves little doubt about the centrality

of relevance.Relevance is not one failure among four. It is the load-bearing one for the sense of uncooperativeness. The other maxims pattern more narrowly, and their signals are weaker. Under-informativeness is resented more than verbosity. Quality splits between high-trust and low-truthfulness findings. Manner registers but stays secondary (Panfili et al., 2021; Aziz, 2025). The relevance result is the single most robust finding in this literature. Any account of machine pragmatic failure must explain it.

5.2 Comprehension failures cluster where unstated context must be reconstructed

Misreadings concentrate in particular task environments. They accumulate in cases where meaning depends on information beyond the literal phrasing. Evidence from Theory-of-mind research indicates competence in explicate mental-state tasks including false-belief reasoning (Kosinski, 2023). It shows weakness on faux pas, where the social knowledge is not in the text (Strachan et al., 2024). Intent recognition is strong on frequent, canonically phrased intents. It is weak in the long tail, where purpose must be inferred rather than matched (Bodonhelyi et al., 2024). Implicature, the recovery of meaning that is meant but unstated, is both the central case and the least systematically measured. Benchmarks are acknowledged to under-cover it (Salman &Matrood, 2025; Ma et al., 2025).The boundary these results draw is consistent. Machine comprehension is best where the pragmatic work is pre-compiled into the form of the input. It degrades as interpretation comes to depend on context the model does not represent. This is the same boundary the faux pas finding draws, between tracked propositional content and situated normative judgement. It identifies the comprehension-side locus of pragmatic failure: the reconstruction of unstated context.

5.3 Two dynamic conditions are consequential and under-characterised

Dialogue systems face two dynamic conditions which remain understudied. The first, pragmatic drift, emerges when cumulative turns alter contextual alignment producing divergence between interlocuters' assumption (Strachan et al., 2024). The second is the opposite case. Ambiguity or under specification compels the system to either elicit clarification or commit to interpretation. Both lack systematic characterization by small-scale tests (Liu et al., 2023).The difficulty grows once the channel is noisy or multimodal. He and Garner(2023) evaluate large language models on spoken input, stating that "its performance is sensitive to ASR errors" (p. 1), it also performs worse on slot filling .It would distinguish the two at the level of definition, not treat warrant failure as an exotic sub-case of recovery failure. The distinction is not merely tidy. It predicts different repairs. Recovery failure is addressed by supplying or modelling context. Warrant failure is addressed by constraining output to what context licenses.Recognition errors upstream of interpretation mean the model reasons over a degraded input (He & Garner, 2023). Vision-language systems show similar fragility. Their answers shift under small, pragmatically irrelevant changes to the image (Shah et al., 2025). The recurring deficit is pragmatic stability: holding a consistent interpretation across turns, and across perturbations that should not matter. That is exactly what these conditions probe, and what the evidence base least covers.

5.4 Sociopragmatic failure is most visible beyond English

Thomas's second category appears mainly where the relevant norms are culturally marked enough to make miscalibration legible. That means largely outside English. Korean evaluation documents over-formality and honorific misuse. The honorific grammar is handled. Its social calibration is not (Park et al., 2024). Arabic evaluation documents literalism, a pragmalinguistic failure of form-to-force mapping. It also documents dialect mixing, a sociopragmatic failure to track the variety the situation calls for (Alshraah et al., 2025)The coding placed these phenomena cleanly on the Thomasian axis. That confirms the framework's reach. It also exposes a blind spot. An English-centred, decontextualised research programme cannot see sociopragmatic failure, because the norms it would violate are not in view (Strachan et al., 2024; Abdurahman et al., 2024). The cross-linguistic studies are where the most culturally consequential category becomes observable. They are also where the evidence is thinnest relative to its importance (Alshraah et al., 2025; Ma et al., 2025).

5.5 Hallucination operates as a failure of warrant

Hallucination falls on the production side of dyad. It resists the recovery-failure reading that fits the comprehension findings. Within speech-act account, hallucination constitutes an assertion that lacks the licensing condition of evidence, authority, situational fit. That makes it an infelicity, not a falsehood (Li, 2025).The felicity reading explains features the factual reading leaves puzzling. Hallucinations do most harm when most fluent. Fluency supplies the surface signals of warranted assertion exactly where the licensing conditions are absent. The harm concentrates in high-stakes domains, where the gap between the performed assertion and its missing warrant matters most (Aditya, 2024). The reading also explains the leading mitigations. Retrieval grounding, chain-of-verification, human-in-the-loop review, citation checking, neurosymbolic fact-checking, and multi-agent critique are best understood as ways to supply the missing warrant and to regulate what the system may assert (Aditya,

2024; Lamba et al., 2025). Even the taxonomies point this way. Sorting hallucinations as input-, context-, or fact-conflicting already grades them by relation to conversational purpose, whatever the surrounding prose says (Ye et al., 2023; Huang et al., 2023). Hallucination thus codes as the sharpest case of failure of warrant. The fault lies not in the content of an assertion but in the conditions under which the output counts as one.

5.6 Pragmatic failure in machine discourse is doubly displaced

The four substantive findings converge on one cross-cutting result about participant structure. Every coded phenomenon satisfies Thomas's test. Well-formed language fails to deliver the meaning the situation required, with grammar intact. Yet each strains the framework at the same joint. Where the user speaks, the misreading hearer is a system whose failures gather where context cannot be read off the surface (Strachan et al., 2024). Where the model is the source, no intention stands behind the utterance for the user's reading to be measured against (Li, 2025).The relevance-theoretic restatement locates failure in the speaker's utterance, the hearer's interpretation, or both, and it assumed both sources were available (Zhang, 2021). In machine talk one is, in each direction, structurally vacant. Machine pragmatic failure is in this sense doubly displaced. It is recognisable by its symptom and unlike its human model in its structure. The displacement is the finding the discussion must interpret.

6. Discussion

6.1 Recovery versus warrant: re-specifying the construct

The five findings necessitate a revision. Pragmatic failure in AI-generated discourse is not a single construct but two distinct kinds. They are joined by symptoms, yet they are divided: recovery belongs comprehension side, warrant to the production side. Recovery failures happen when the model misreads a genuinely intendedhuman meaning. These errors are continuous withinterlanguage pragmatics failure, where learners misinterpret or misapply pragmatic rules. Thomas’s distinction between pragmalinguistic and sociopragmaticfailureapplies here. Literalism, honorific miscalibration, and dialect mismatch exemplify this continuity (Alshraah et al., 2025; Park et al., 2024).Failures of warrant are of a different kind. In these cases, the model produces text whose illocutionary force exceeds anything it is positioned to mean or commit to. Unlike recovery errors this has no clear analogue to human discourse. Human speakers do not ordinarily generate the full surface apparatus of warranted assertion without any backing in evidence or situational fit. The closest classical neighbor is infelicity, not mismatch. For this reason, the speech-act theory provides a sharper instrument to analyze warrant failure (Li, 2025).A re-specified construct makes the division central. It defines recovery and warrants as distinct form of pragmatic failure, rather than treating warrant as a sub-case of recovery. The distinctive is substantive because it highlights different repair strategies. Recovery failure is addressed by supplying or modelling the missing context. Warrant failure is repaired by constraining output to what context licenses.

6.2 Why extending Grice is more than housekeeping

The proposals to supplement Grice's maxims do principled work, read against these findings. The maxims of benevolence and transparency specify, for a partner that cannot be presumed cooperative because it cannot be presumed to intend, properties that stand in for the missing presumption (Miehling et al., 2024). That is the right move once the partner's mind is the thing that has gone missing. It converts an assumption about the partner into a requirement on the system.Treating Gricean norms as design targets performs the same conversion from the other direction. It builds the cooperative behaviour that can no longer be presumed. Agents equipped with the norms come out clearer, more accurate, and more relevant than untutored ones (Murukannaiah & Singh, 2025; Kim et al., 2025). The speech-act gating of assertion performs the conversion for warrant. It replaces the assumption that an asserter has evidence with a check on whether the conditions for assertion hold (Li, 2025). The common move across these proposals is to externalise, as an engineered property or an explicit check, what the classical framework took for granted in a human partner. The findings explain why the move is needed. The presumptions the maxims rely on are exactly what a non-intending interlocutor cannot supply.

6.3 From two parties to a coupled process

The displacement finding in Section 5.6, and the distributed-cognition reading of hallucination (Osler, 2026), suggest a shift in the unit of analysis. If one party does not interpret in the relevant sense, then pragmatic success or failure is realised somewhere other than in that party's mind. It is realised in the arrangement around it: the prompt, the retrieved context, the verification layer, the human reading.This converges with the pressure cross-cultural pragmatists already named, where heterogeneous lingua-franca norms unsettled the picture of a stable speaker meaning recovered against a stable cultural frame (Peter, 2019). A re-specified account would treat pragmatic failure less as a discrete event in one party's error. It would treat it more as a breakdown in a jointly held interpretive process, one of whose participants does not intend. Reading the exchange as genuinely joint

sense-making, rather than as a test the machine passes or fails, reframes the design question. The question is no longer how to make the model cooperative. It is how to make the coupled system cooperative (Rappa et al., 2026).

6.4 Implications for design and pedagogy

The analysis yields its practical implications rather than decorating them. The model's failures gather where unstated context must be reconstructed (Section 5.2) and where output runs ahead of warrant (Section 5.5). The interventions that help are therefore the ones that supply context or constrain output to what context licenses. Retrieval grounding and verification address the warrant problem (Aditya, 2024; Lamba et al., 2025). Explicit norm modelling addresses the cooperation problem (Murukannaiah & Singh, 2025). Benchmarks that actually probe implicature, reference, and cross-cultural appropriateness address the measurement problem (Ma et al., 2025; Alshraah et al., 2025).The educational literature adds a human-side corollary. As these systems mediate more communication, users' own pragmatic competence is implicated. It is both at risk and potentially supported, depending on how the interaction is structured (Eragamreddy, 2025). The metaphor through which the failure is described carries its own consequences. Calling errors "hallucinations" shapes expectations, blame, and attributed agency. That is itself a site of pragmatic effect at the institutional level (Förster & Skop, 2025).

7. Conclusion

The analysis confirms that pragmatic failure survives the move from human conversation to human-AI interaction, but not without modification. Thams's concept now describes a system that rarely violates grammar but frequently fails in context- sensitive communication (Thomas, 1983; Mahowald et al., 2024). The model remains useful at the analytical level.Its value is evident in thethree recuring patterns: failure of relevance, difficulty in reconstructing unstated context, and unwarranted assertions(Panfili et al., 2021; Nam et al., 2023; Strachan et al., 2024; Bodonhelyi et al., 2024; Li, 2025). At the same time, the model reaches its limits because it presupposes two intending interlocuters, a condition that AI-mediatedinteraction no longer satisfies. The evidence of this study supports revisions rather than replacement. The revised category separates failures of recovery from failures of warrant. The coupled human-machine process, not the machine alone, should be the unit in which pragmatic success or failure is realised. The call for such a redefinition predates the present systems (Peter, 2019). Their arrival supplied both the urgency and the test cases.The limitation of this paper arises from the available evidence. The synthesis reflects a literature that is mostly English-language, text-based, and concentrated on a few systems. Its coding is interpretive where sources did not use the applied categories. The agenda these limits set is concrete. The field needs cross-linguistic, multimodal, and multi-turn evaluation of how meaning is reconstructed and warranted in real interaction, supported by shared, transparent benchmarks, in place of the controlled, text-only, English-centred testing on which most present claims rest (Abdurahman et al., 2024; Ma et al., 2025). Until then, the field will keep recognising pragmatic failure in AI-generated discourse more readily than it can measure it. The next challenge is not recognizing pragmatic failure but establishing firmer grounds for claiming it.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. PNAS Nexus, 3. https://doi.org/10.1093/pnasnexus/pgae245

Aditya, G. (2024). Understanding and addressing AI hallucinations in healthcare and life sciences. International Journal of Health Sciences. https://doi.org/10.47941/ijhs.1862

Alshraah, S. M., Aldaghri, A., & Oraif, I. (2025). AI's struggle with Arabic: A study on pragmatic failures in contextual communication. Forum for Linguistic Studies. https://doi.org/10.30564/fls.v7i11.9252

Aziz, A. L. A. A. (2025). AI and pragmatics: Do chatbots follow speech acts and maxims? Evaluating AI-generated conversations using pragmatics. Wasit Journal for Human Sciences. https://doi.org/10.31185/wjfh.vol21.iss3.1012

Bodonhelyi, A., Bozkir, E., Yang, S., Kasneci, E., & Kasneci, G. (2024). User intent recognition and satisfaction with large language models: A user study with ChatGPT. arXiv. https://doi.org/10.48550/arxiv.2402.02136

Eragamreddy, N. (2025). The impact of AI on pragmatic competence. Journal of Teaching English for Specific and Academic Purposes. https://doi.org/10.22190/jtesap250122015e

Firdaus, T., Bakari, L. Y., Octarianto, D. S., Khoirunnisa, R., Malini, D. R., & Mukarromah, M. (2025). Applying Grice's maxims and speech act theory to analyze pragmatics in ChatGPT and Gemini AI dialogues. International Journal of Language and Translation Studies. https://doi.org/10.63673/lotus.1631745

Förster, S., & Skop, Y. (2025). Between fact and fairy: Tracing the hallucination metaphor in AI discourse. AI and Society. https://doi.org/10.1007/s00146-025-02392-w

Hamid, O. H. (2024). Beyond probabilities: Unveiling the delicate dance of large language models (LLMs) and AI-hallucination. In 2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA) (pp. 85–90). https://doi.org/10.1109/cogsima61085.2024.10553755
He, M., & Garner, P. N. (2023). Can ChatGPT detect intent? Evaluating large language models for spoken language understanding. arXiv. https://doi.org/10.48550/arxiv.2305.13512
Huang, L., Yu, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43, 1–55. https://doi.org/10.1145/3703155
Jacquet, B., Hullin, A., Baratgin, J., & Jamet, F. (2019). The impact of the Gricean maxims of quality, quantity and manner in chatbots. In 2019 International Conference on Information and Digital Technologies (IDT) (pp. 180–189). https://doi.org/10.1109/dt.2019.8813473
Kim, Y., Chin, B., Son, K., Kim, S., & Kim, J. (2025). Applying the Gricean maxims to a human-LLM interaction cycle: Design insights from a participatory approach. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3706599.3719759
Kosinski, M. (2023). Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences, 121. https://doi.org/10.1073/pnas.2405460121
Krause, L., & Vossen, P. (2024). The Gricean maxims in NLP: A survey (pp. 470–485). https://doi.org/10.18653/v1/2024.inlg-main.39
Lamba, N., Tiwari, S., & Gaur, M. (2025). Hallucinations in scholarly LLMs. Open Conference Proceedings. https://doi.org/10.52825/ocp.v8i.3175
Li, A. (2025). Hallucination as pragmatic failure: A theoretical reframing of large language models. The Journal of Interactive Social Sciences. https://doi.org/10.64744/tjiss.2025.38
Liu, Y.-H., Han, T., Zhang, J.-Y., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology. https://doi.org/10.1016/j.metrad.2023.100017
Ma, B., Li, Y., Zhou, W., Gong, Z., Liu, Y. J., Jasinskaja, K., Friedrich, A., Hirschberg, J., Kreuter, F., & Plank, B. (2025). Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges [Preprint]. arXiv. https://arxiv.org/abs/2502.12378
Mahowald, K., Ivanova, A. A., Blank, I., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. Trends in Cognitive Sciences, 28, 517–540. https://doi.org/10.1016/j.tics.2024.01.011
Miehling, E., Nagireddy, M., Sattigeri, P., Daly, E. M., Piorkowski, D., & Richards, J. T. (2024). Language models in dialogue: Conversational maxims for human-AI interactions. arXiv. https://doi.org/10.48550/arxiv.2403.15115
Murukannaiah, P. K., & Singh, M. P. (2025). Gricean norms as a basis for effective collaboration (pp. 1812–1820). https://doi.org/10.48550/arxiv.2503.14484
Nam, Y., Chung, H., & Hong, U. (2023). Language artificial intelligences' communicative performance quantified through the Gricean conversation theory. Cyberpsychology, Behavior, and Social Networking, 26, 919–923. https://doi.org/10.1089/cyber.2022.0356
Osler, L. (2026). Hallucinating with AI: Distributed delusions and "AI psychosis." Philosophy and Technology, 39. https://doi.org/10.1007/s13347-026-01034-3
Panfili, L., Duman, S., Nave, A., Ridgeway, K., Eversole, N., & Sarikaya, R. (2021). Human-AI interactions through a Gricean lens. arXiv. https://doi.org/10.3765/plsa.v6i1.4971
Park, D., Lee, J., Jeong, H., Park, S., & Lee, S. (2024). Pragmatic competence evaluation of large language models for the Korean language (pp. 256–266).
Peter, M. (2019). Cross-cultural pragmatic failure. Training Language and Culture. https://doi.org/10.29366/2019tlc.3.1.5
Quan, Z., & Chen, Z. (2024). Human-computer pragmatics trialled: Some (im)polite interactions with ChatGPT 4.0 and the ensuing implications. Interactive Learning Environments, 33, 1020–1039. https://doi.org/10.1080/10494820.2024.2362829
Rappa, N., Tang, K., & Cooper, G. (2026). Making sense together: Human-AI communication through a Gricean lens. Linguistics and Education. https://doi.org/10.1016/j.linged.2025.101489
Ren, Y., Li, Y., & Wang, Q. (2016). Manifestation of pragmatic failure in intercultural communication (pp. 1039–1044). https://doi.org/10.2991/icemaess-15.2016.214

Salman, Y., & Matrood, D. (2025). Conversational implicature in human-AI interactions. Frontiers in Global Research. https://doi.org/10.55559/fgr.v1i3.22
Saygin, A. P., & Çiçekli, I. (2002). Pragmatics in human-computer conversations. Journal of Pragmatics, 34, 227–258. https://doi.org/10.1016/s0378-2166(02)80001-7
Shah, M., Balaji, S., Sarkhel, S., Dey, S., & Venugopal, D. (2025). Analyzing the sensitivity of vision language models in visual question answering. arXiv. https://doi.org/10.48550/arxiv.2507.21335
Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 8, 1285–1295. https://doi.org/10.1038/s41562-024-01882-z
Thomas, Jenny A. (1983). Cross-cultural pragmatic failure. Applied Linguistics, 4, 91–112. https://doi.org/10.1093/applin/4.2.91
Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. arXiv. https://doi.org/10.48550/arxiv.2309.06794
Zhang, J. (2021). A tentative analysis of pragmatic failure from the perspective of relevance theory. The Frontiers of Society, Science and Technology. https://doi.org/10.25236/fsst.2021.030306