COGENT: Multiagent Large Language Models Communicate with self-coined terms

Anonymous ACL submission

Abstract

A central bottleneck in LLM-based multi-agent systems is not whether agents can talk, but what their communication is made of. Most protocols exchange ordinary natural-language traces in fixed answer–critique formats, forcing recurring proof states, option contrasts, and variable relations to be repeatedly rephrased. We introduce COGENT (Collective Grounding of Emergent Named Terms), a black-box communication layer that augments trace-only exchange with a coin–share–use lifecycle. Agents coin task-local terms with explicit definitions and same-turn use, share accepted entries through a shared term board, and then reuse, challenge, or repair them in later interaction. On pilot slices of MMLU, FOLIO, and GSM8K, COGENT improves accuracy over vanilla debate on five of six model–task pairs and reduces post-warm-up debate tokens across all evaluated pairs. In a five-game Werewolf stress test, public terms also receive far stronger reuse and cross-speaker uptake than faction-private terms. These results suggest that MAS communication can improve by redesigning its basic unit: from longer traces to shared task-local words that agents can use together.

1 Introduction

In LLM-based multi-agent systems (MASs), a key bottleneck is not only how many agents speak, but what communicative objects they share (Wooldridge, 2009; Dorri et al., 2018; Guo et al., 2024; Li et al., 2024; Han et al., 2024a; Yan et al., 2025). Existing MAS protocols let agents compare answers, critique claims, and revise decisions, usually by passing natural-language traces through fixed answer–critique formats (Du et al., 2024; Madaan et al., 2023; Shinn et al., 2023; Li et al., 2023; Wu et al., 2024; Hong et al., 2024; Qian et al., 2024). Such traces are expressive, but weak as reusable infrastructure: the same proof status, option contrast, or variable relation must be rephrased across turns, which can lengthen interaction or drift in meaning.

This paper studies lexical innovation as a communication primitive for LLM-based MASs. Human teams stabilize collaboration by inventing local names: a shorthand for a case distinction, a nickname for a failure mode, or a variable name for a recurring quantity (Clark and Marshall, 1981; Lewis, 1969; Brennan and Clark, 1996). The key property of such names is not novelty, but uptake. A local expression becomes useful only when others can apply, challenge, or repair it under the same task criterion. We ask whether LLM agents can similarly coin task-local terms, expose them to collaborators, and use them as shared reasoning units.

We introduce COGENT, *Collective Grounding of Emergent Named Terms*, a black-box communication layer built around a coin–share–use lifecycle. In *Coin*, each agent may create a term in the form `term = brief definition`, accepted only if it is used in the same reasoning turn. In *Share*, accepted entries are placed in a shared `KNOWN_TERMS` block. In *Use*, later interaction is conditioned on this board, allowing agents to reuse, contest, or repair coined terms rather than repeatedly reconstruct the same reasoning structure. Figure 1 illustrates the lifecycle.

The design follows a use-based account of meaning: a coined term is not grounded because it is short, but because it supports rule-governed public use (Wittgenstein, 1953; Grice, 1975; Clark and Marshall, 1981; Clark and Brennan, 1991; Brennan and Clark, 1996; Lewis, 1969; Pickering and Garrod, 2004). COGENT therefore treats term creation, visibility, and uptake as distinct events. Coining proposes a lexical handle; sharing makes it available; use tests whether it preserves the task criterion.

We evaluate COGENT on MMLU, FOLIO, and GSM8K slices, with an exploratory Werewolf hidden-information setting. COGENT improves

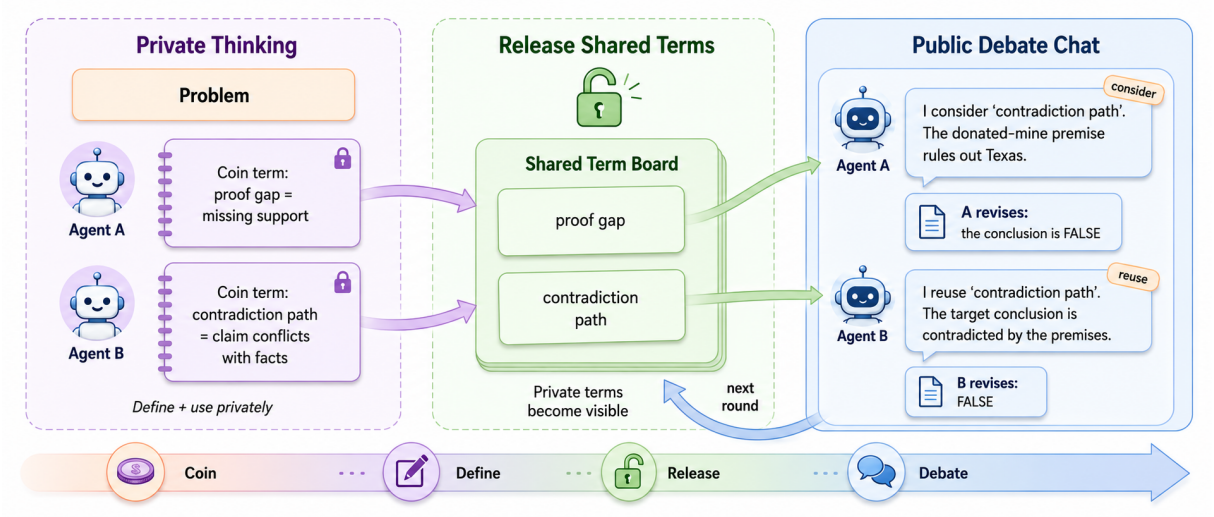

Figure 1: Overview of the COGENT coin–share–use communication layer. Coin creates definition-bound lexical candidates; Share exposes accepted entries; Use tests public uptake in later interaction.

accuracy over vanilla debate on five of six model–task pairs and reduces post-warm-up debate tokens across all evaluated pairs; Werewolf further shows stronger reuse and cross-speaker uptake for public terms than faction-private terms. Overall, we shift MAS design from exchanging more traces to creating shared task-local words that make later communication more stable, auditable, and efficient.

2 Related Work

Communication protocols in LLM-based multi-agent systems. Classical MAS research treated communication and coordination as protocol-level design problems, from contract nets to broader multi-agent architectures (Smith, 1980; Wooldridge, 2009; Dorri et al., 2018). Recent LLM reasoning systems increasingly treat inference as interaction rather than a single forward pass. Chain-of-thought prompting elicits intermediate rationales (Wei et al., 2022); self-consistency samples and aggregates multiple reasoning paths (Wang et al., 2023); Tree-of-Thoughts searches over partial solution states (Yao et al., 2023); and Self-Refine and Reflexion use verbal feedback to iteratively revise outputs (Madaan et al., 2023; Shinn et al., 2023). Multi-agent debate assigns several model instances to propose answers, inspect one another's claims, and update over rounds (Du et al., 2024). Broader LLM-agent frameworks, social-simulation studies, and surveys likewise foreground conversation, role structure, workflow orchestration, and collaboration channels as central design choices (Li et al., 2023; Park et al., 2023; Wu et al., 2024; Hong et al., 2024; Qian et al., 2024; Gao et al., 2024; Dai et al., 2026; Zou et al., 2025; Guo et al., 2024; Li et al., 2024; Han et al., 2024a; Yan et al., 2025). These systems show that communication can improve reasoning, but usually treat the medium as transparent: agents pass ordinary natural-language traces through fixed answer–critique formats. COGENT asks whether the communication substrate itself can be improved by allowing agents to create task-local lexical objects, share them, and test their public use.

Compression, memory, and reusable reasoning state. Long deliberation can preserve intermediate information, but also increases context length and makes later turns depend on verbose, paraphrased, or unstable prior text (Shannon, 1948; Zipf, 1949). Summarization, memory-style prompting, and feedback histories can reduce surface length or retain critiques (Madaan et al., 2023; Shinn et al., 2023; Zou et al., 2025). However, a summary is still typically a restatement of earlier text. It may be shorter without becoming a reusable object of agreement, disagreement, or repair. COGENT separates lexical reuse from ordinary compression. A candidate term must be introduced with an explicit definition, used in the same reasoning turn, exposed through a shared term board, and evaluated by later uptake. This makes the communication unit auditable: a term can be compact but ungrounded if no other agent can use it consistently, while a longer coined expression can still preserve the task criterion needed for later reasoning.

Grounding, conceptual pacts, and lexical coordination. COGENT also connects LLM-MAS design to theories of dialogue grounding and lexical coordination. Work on semantic coordination, conceptual pacts, and interactive alignment shows that communication is not completed by producing an utterance; interlocutors establish local conventions shaped by the current joint activity (Grice, 1975; Lewis, 1969; Clark and Marshall, 1981; Garrod and Anderson, 1987; Clark and Brennan, 1991; Brennan and Clark, 1996; Pickering and Garrod, 2004; Garrod and Pickering, 2004). Emergent-communication work studies learned symbol systems under multi-agent rewards (Foerster et al., 2016; Lazaridou et al., 2017; Havrylov and Titov, 2017; Mordatch and Abbeel, 2018); COGENT instead exposes definition-bound terms inside ordinary LLM interaction. A local name becomes efficient only after participants know how to use it, when to reject it, and what distinction it preserves. COGENT operationalizes this insight for LLM-based MASs. Its coined terms are not general dictionary entries, nor are they evaluated by novelty alone. They are task-local lexical commitments whose value depends on provenance, visibility, cross-speaker uptake, reuse, repair, and downstream decision effects. In this sense, COGENT reframes multi-agent communication from exchanging more text to building shared words that structure later reasoning.

3 Method: The Coin–Share–Use Communication Layer

COGENT is a black-box communication layer around a base LLM f_θ (Vaswani et al., 2017). For each problem q, it instantiates two agents, AgentA and AgentB, with independent histories, agent-private self-coined lexicons L_A and L_B, and a shared term board L^{pub} visible to both agents. The protocol follows a coin–share–use lifecycle. Agents first solve independently to preserve the same starting condition as vanilla debate; they then coin task-local terms, share accepted entries, and use the shared terms in later lexicon-conditioned interaction. Figure 1 gives the compact view, while Algorithm 1 formalizes the experimental control logic.

Pre-step: ordinary independent solving. Each agent first answers independently:

$$m_i^0 = f_\theta(P_{\text{init}}(q)), \qquad i \in \{A, B\}.$$

Algorithm 1 COGENT Coin–Share–Use Communication$(q; R, K, R_c)$

Input: instance q, meta-rounds R, coining passes K, active coining rounds R_c
Output: final answer $\hat{a}$

$L_A, L_B, L^{\text{pub}} \leftarrow \emptyset$; initialize histories
Pre-step: ordinary independent solving
for $i \in \{A, B\}$: $m_i^0 \leftarrow \text{LLMStep}(i, P_{\text{init}}(q))$
$\ell_A, \ell_B \leftarrow m_A^0, m_B^0$
for $r = 1$ to R **do**
  $K_r \leftarrow K$ if $r \le R_c$, otherwise 0
  Coin
  for all $i \in \{A, B\}$ **do**
    $(L_i, \ell_i) \leftarrow \text{Coin}(i, \ell_i, L_i, L^{\text{pub}}, K_r)$
  end for
  Share
  $p_A, p_B \leftarrow \ell_A, \ell_B$; $L^{\text{pub}} \leftarrow L_A \cup L_B$
  Use
  $m_A^r \leftarrow \text{Use}(A, q, p_B, L^{\text{pub}})$; update L_A, ℓ_A
  $L^{\text{pub}} \leftarrow L_A \cup L_B$
  $m_B^r \leftarrow \text{Use}(B, q, m_A^r, L^{\text{pub}})$; update L_B, ℓ_B
end for
$\hat{a} \leftarrow \text{Finalize}(m_A^R, m_B^R)$

This produces the same starting condition as vanilla sequential debate, with no lexicon field in the initial prompt. The pre-step is not part of the lexical lifecycle; it ensures that COGENT and vanilla debate begin from comparable ordinary answers.

Step 1: Coin. During the first R_c meta-rounds, each agent performs K coining passes over its own latest response:

$$(\ell_i^{r,k}, \Delta L_i^{r,k}) = f_\theta\big(P_{\text{name}}(\ell_i^{r,k-1}, L_i^{r,k-1} \cup L_{r-1}^{\text{pub}})\big).$$

During Coin, each agent revisits its latest reasoning and constructs an agent-private self-coined lexicon. The prompt uses a protocol-defined term-definition format: a candidate is accepted only if it appears as $t = d$, is used in the same reasoning trace, and passes novelty/blocklist checks. It also forbids final-answer markers so that coinage does not become a hidden answer-selection step. This implements a use-bound definition: a term must already do inferential work for the agent that coined it. Definition gives the term a local binding, while same-turn use tests whether it functions in the reasoning trace rather than serving as a decorative label. Thus Coin is not ordinary summary compression: a summary can shorten a trace without creating an object of later agreement, disagreement, repair, or replacement.

Step 2: Share. At communication boundaries, accepted entries from the agent-private lexicons

are shared through the shared term board:

$$L_r^{\text{pub}} = L_A^r \cup L_B^r.$$

Each shared entry contains the surface term, definition, and provenance, making it available as an explicit communication resource rather than an implicit phrase buried in a trace. Share is a visibility transition, not grounding itself: it exposes a coined term to collaborators, but later interaction must still test whether others can use it under the intended task criterion.

Step 3: Use. Use is implemented as lexicon-conditioned sequential debate. The shared term board is injected into the public interaction prompt, and agents may reuse, challenge, repair, or ignore available coined terms. The sequential updates are

$$m_A^r = f_\theta\Big(P_{\text{debate}}(q, p_B^r, L_r^{\text{pub}})\Big),$$
$$m_B^r = f_\theta\Big(P_{\text{debate}}(q, m_A^r, L_r^{\text{pub}})\Big).$$

Use tests coined terms as public reasoning objects. A term that is shared but never taken up remains a private label; a term reused, challenged, or repaired around the same task criterion becomes evidence of public utilization.

Post-step: finalization. The controller extracts predictions only from the latest interaction responses and returns

$$\hat{a} = \textsc{Finalize}\big(\text{Extract}(m_A^R), \text{Extract}(m_B^R)\big).$$

Finalization is kept separate from the coin–share–use lifecycle: it reads the latest predictions, while the lifecycle describes how agents build and use the communication layer.

Metrics. We report task accuracy, compression rate, and post-warm-up token count. For COGENT, if B_n^{cog} is the sum of the agents' initial-output lengths and W_n^{cog} is the sum of their post-warm-up private-state lengths, then

$$\text{CR}_n^{\text{cog}} = \frac{W_n^{\text{cog}}}{B_n^{\text{cog}}}.$$

The token column for COGENT reports the average length of the final no-compression debate round after warm-up, not total API usage. For Werewolf, where there is no single gold answer, we additionally report lexical uptake:

$$\text{Uptake} = \frac{N_{\text{reuse}}}{N_{\text{term}}},$$

where N_{reuse} is the number of observed term reuses and N_{term} is the number of accepted terms.

4 Experimental Setup

We evaluate COGENT on three answer-oriented benchmarks: MMLU for broad multiple-choice knowledge (Hendrycks et al., 2021), FOLIO for natural-language reasoning with first-order logic annotations (Han et al., 2024b), and GSM8K for grade-school mathematical reasoning (Cobbe et al., 2021). We use cost-controlled slices of 500 MMLU examples, 282 FOLIO examples, and 500 GSM8K examples.

The main experiments use GPT-3.5-Turbo and Qwen3-35B-A3B Instruct with thinking mode disabled. For each model–benchmark pair, we compare three variants: SINGLE, a one-agent answer; VANILLA, two-agent sequential debate without a lexicon channel; and COGENT, which adds the coin–share–use communication layer. Unless otherwise stated, COGENT uses $(R, K, R_c) = (3, 2, 2)$: three meta-rounds, two coining passes per active warm-up round, and two coining-active rounds before debate-only interaction.

For lexical-function analysis, benchmark terms are softly annotated into six categories: discriminative cue, domain concept, quantity/unit variable, operation checkpoint, proof-status logic, and gap/uncertainty marker. We use human annotation with LLM-assisted verification and manual adjudication; Appendix C gives the full protocol. We also include a five-game Werewolf stress test. Because Werewolf produces full hidden-role game trajectories rather than standardized answer markers, it is analyzed separately through lexical-grounding metrics. In this setting, COGENT maintains both a public term pool visible to all players and a wolf-private pool visible only to the wolf faction.

5 Results

Table 1 tests whether the coin–share–use communication layer improves the accuracy–efficiency profile of multi-agent interaction. It shows three main patterns. First, COGENT improves accuracy over vanilla debate on five of six model–task pairs. For GPT-3.5-Turbo, accuracy rises from 64.80% to 65.20% on MMLU, from 40.20% to 47.06% on FOLIO, and from 78.20% to 79.40% on GSM8K. For Qwen3-35B-A3B, it rises from 87.33% to 87.80% on MMLU and from 85.78% to 87.25% on FOLIO. The only regression is Qwen3-35B-A3B on GSM8K, where COGENT reaches 96.00% compared with 96.59% for vanilla debate.

Second, the strongest gains appear on FOLIO.

Model	Variant	MMLU (500)			FOLIO (282)			GSM8K (500)		
		Acc.	CR	Tok.	Acc.	CR	Tok.	Acc.	CR	Tok.
GPT-3.5-Turbo	Single	60.00	N/A	127.02	40.69	N/A	**105.06**	67.20	N/A	89.34
	Vanilla	64.80	0.91	209.98	40.20	0.99	195.74	78.20	1.97	99.22
	COGENT	**65.20**	**0.40**	**100.17**	**47.06**	**0.46**	120.74	**79.40**	**0.79**	**86.47**
Qwen3-35B-A3B w/o think	Single	86.13	N/A	435.49	81.86	N/A	450.80	96.29	N/A	**99.63**
	Vanilla	87.33	1.39	1031.98	85.78	1.09	530.42	**96.59**	1.45	373.70
	COGENT	**87.80**	**0.42**	**265.47**	**87.25**	**0.42**	**344.49**	96.00	**0.70**	190.18

Table 1: Main results on MMLU, FOLIO, and GSM8K. Accuracy is reported in percentage. *CR* is the compression rate, where lower values indicate stronger warm-up shortening. *Tok.* denotes raw single-answer tokens for SINGLE and post-warm-up debate tokens for VANILLA/COGENT, not total API usage.

This is consistent with the role of public terms: FOLIO requires agents to preserve entailment, contradiction, and underdetermination across turns. A well-defined public term can stabilize proof status and prevent later debate from rephrasing a premise relation incorrectly.

Third, COGENT consistently reduces post-warm-up debate tokens relative to vanilla debate. The reductions are approximately 52.3%, 38.3%, and 12.9% for GPT-3.5-Turbo on MMLU, FOLIO, and GSM8K, respectively. For Qwen3-35B-A3B, the corresponding reductions are approximately 74.3%, 35.1%, and 49.1%. These token numbers should be read as downstream debate length after warm-up, not total API cost. Together, the results support a cautious interpretation: COGENT is a communication protocol whose success depends on whether shared coined terms preserve the criteria needed for the final answer. The Qwen3-35B-A3B GSM8K result shows that stronger shortening can coexist with a small loss in exact arithmetic accuracy.

Exploratory Werewolf summary. Table 2 gives a compact summary of the Werewolf stress test. Across five games, COGENT accepts 32 terms and observes 417 reuses, yielding a lexical uptake rate of 13.03. Public exposure is especially productive: public terms represent 37.5% of accepted entries but account for 69.5% of observed reuse, with a reuse-per-entry rate 3.81 times that of wolf-private terms. These numbers should not be read as a clean full-game efficiency win; rather, they show that COGENT makes convention formation measurable under hidden information.

6 Analysis

6.1 From trace exchange to coined-term use

The useful abstraction for COGENT is not compression alone, but criterion-preserving use. A compression analysis asks whether a coined string shortens a trace. A communication analysis asks what counts as using that string correctly within the local activity (Wittgenstein, 1953). On this view, a term matters only if it preserves the task criterion that later agents must continue to apply.

The benchmark lexicons make this task-dependence visible. MMLU turns coinage into classificatory distinction, GSM8K into local mathematical notation, FOLIO into proof-status tracking, and Werewolf into social-epistemic rule formation under hidden information. The same operation—coining a term—therefore acquires different linguistic functions depending on what the task requires agents to preserve.

6.2 Lexical ecology across tasks

The lexical ecology varies systematically by task; Appendix D reports detailed counts, criteria, and examples. The main finding is not lexicon size alone, but functional specialization. In MMLU, useful terms name option-level contrasts; in GSM8K, they behave like temporary variables or operation labels; in FOLIO, they mark entailment, contradiction, or underdetermination; and in Werewolf, they name public stances, pressure sets, role-claim effects, and voting responsibilities. This supports the central thesis of COGENT: public meaning is a relation among term, definition, task norm, uptake, and downstream judgment (Lewis, 1969; Clark and Brennan, 1991).

Figures 2 and 3 make the same point quanti-

Metric	Value	Interpretation
Accepted Werewolf terms	32	A compact task-local lexicon is formed across five hidden-role games.
Observed term reuses	417	Accepted terms are taken up repeatedly rather than remaining one-off labels.
Cross-speaker public uptake	15 / 32 terms	Nearly half of accepted terms are used by at least two public speakers, indicating uptake beyond the inventor.
Public reuse share	69.5%	Public terms contribute most observed reuse despite representing only 37.5% of accepted entries.
Common-prefix token reduction	5.23% public; 4.95% personal	At matched checkpoints, public-log and all-player personal-log contexts are shorter than in ORIGIN.

Table 2: Compact main-text Werewolf lexical-grounding metrics. Unlike Table 1, these metrics evaluate convention formation rather than benchmark accuracy. The full descriptive metric table is reported in Appendix E.

tatively. MMLU terms concentrate on domain concepts and discriminative cues, GSM8K terms on quantity/unit variables and operation checkpoints, and FOLIO terms on proof-status logic and gap/uncertainty markers. Across tasks, both model families produce mixed lexical functions, but Qwen3-35B-A3B shows a stronger proof-status component while GPT-3.5-Turbo leans more toward domain and quantity terms. The surface form of a coined term is therefore less important than the task norm it helps maintain.

6.3 MMLU and GSM8K: classification vs. notation

MMLU turns lexical innovation into a problem of classification. A multiple-choice answer is correct because the agent preserves the contrast separating the best option from plausible distractors. Terms such as `expansiovision`, `treaty-fulfill mandate`, `focus-exclude rule`, and `cue-power fading` function as local discriminative handles. They resemble conceptual pacts in dialogue (Brennan and Clark, 1996), but the pact is inferential rather than referential: a public MMLU term succeeds when later agents can carry forward the same option-level contrast without reconstructing the full comparison.

GSM8K imposes a different criterion: calculational auditability. Terms such as `commissionalize`, `speedtime`, `unitcost`, `pumpcost`, and `josh cookies` behave like task-local mathematical notation, preserving quantities, units, variable roles, and operation checkpoints. The Qwen3-35B-A3B GSM8K result marks the boundary of the mechanism. COGENT substantially shortens downstream debate, but names that compress away exact arithmetic can weaken verification. GSM8K therefore separates notation from abbreviation: good notation shortens by preserving checkable structure; bad abbreviation hides the step another agent must inspect.

6.4 FOLIO: rule-following and proof-status repair

FOLIO is the clearest test of rule-governed lexical use. A correct answer depends on whether the conclusion is entailed, contradicted, or left open by the premises. Useful coined terms are therefore proof-status operators rather than content summaries. `open-model gap` marks underdetermination, `must-follow path` marks necessity, and `contradiction path`, `linkvoid`, and `lacksupport` mark conflict or missing support. Their meaning depends on whether later reasoning continues under the same rule: underdetermination must not become plausibility, and contradiction must preserve the polarity of the target conclusion.

Figure 4 illustrates why the FOLIO gains in Table 1 are not reducible to compression. The Picuris Mountains case requires agents to rule out the Texas branch, infer that the mine is in a New Mexico mountain range, and mark the conclusion as FALSE. Raw answering and vanilla debate fail because the proof polarity is not stabilized. COGENT succeeds because AgentB coins `novsure` for a premise-contradicting conclusion, AgentA takes it up, and AgentB later reinforces the same function with `novcontra`. The decisive move is not the discovery of a new fact, but the creation of a public proof-status operation that another agent can reuse.

6.5 Werewolf: asymmetric common ground

Werewolf extends COGENT from answer-oriented reasoning to social-epistemic coordination under hidden information (Park et al., 2023; Dai et al., 2026). In MMLU, terms preserve option contrasts; in GSM8K, calculational states; in FOLIO, proof statuses. In Werewolf, the relevant criterion is in-

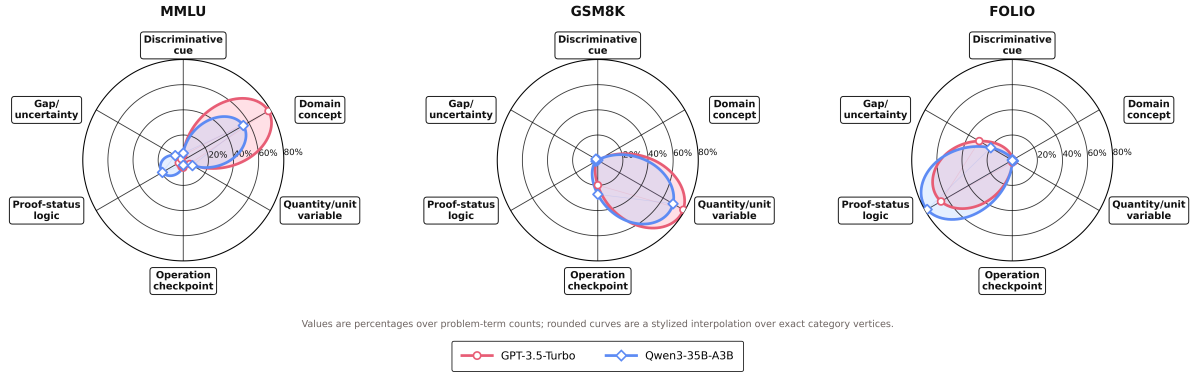

Figure 2: Soft lexical-function radar by task for GPT-3.5-Turbo and Qwen3-35B-A3B. Percentages are computed over problem-term counts after human annotation, LLM-assisted verification, and manual adjudication.

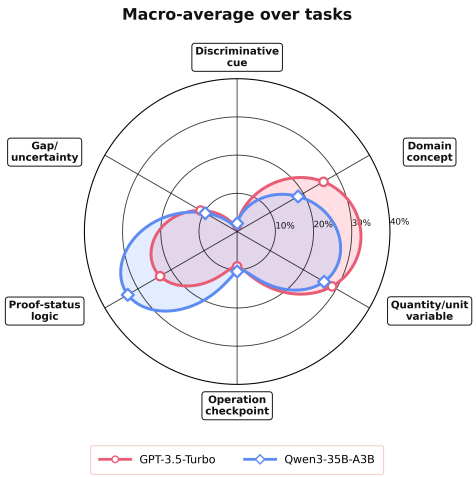

Figure 3: Macro-average soft lexical-function distribution over MMLU, FOLIO, and GSM8K.

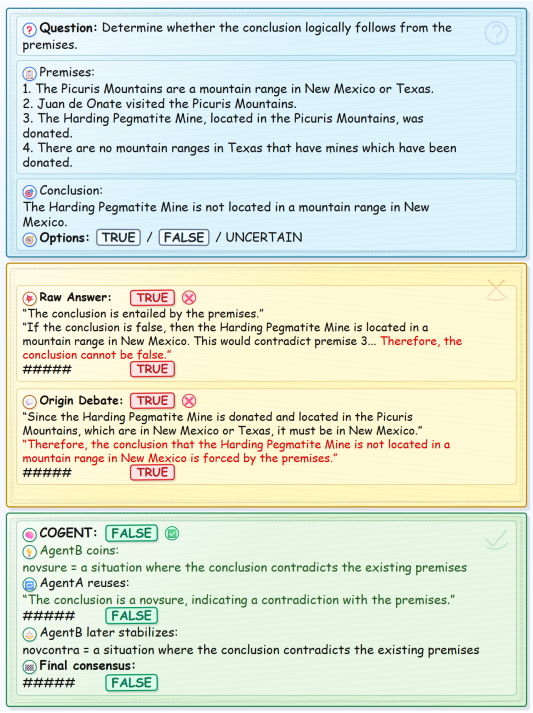

Figure 4: FOLIO keynote case study. Raw answering and vanilla debate predict TRUE, while COGENT predicts FALSE by stabilizing contradiction-status terms such as `novsure` and `novcontra`.

teractional: a term must preserve provisional trust, role-claim evidence, unresolved pressure sets, and voting responsibility. A Werewolf term is meaningful when later agents know when to apply it, when to reject it, and what social move it licenses.

Figure 5 shows four forms of public uptake. `baseline day` becomes a contested early-day frame; its later rejection shows that agents understand the term well enough to dispute its scope. `pressure pool` functions as a shared reference pact for unresolved suspects. `core lock` marks a late-game trust-alignment rule, while `claim gravity` names how a role claim can pull the table toward accepting both the claim and the claimant's reads. These terms become sites of agreement, disagreement, and repair rather than disposable labels.

Table 2 shows the productivity of this lexical layer. The important signal is not vocabulary size alone, but compact vocabulary plus repeated uptake: 32 accepted terms support 417 observed reuses, and public terms are far more productive than faction-private terms. If meaning were merely a private association between an agent and a label, wolf-private entries would be equally strong evidence. Instead, public exposure sharply increases lexical productivity, suggesting that terms placed in the shared day-discussion arena become more reusable than narrower faction-private coinages.

6.6 From coinage to public meaning

Across tasks, COGENT lexicons are audit trails of attempted norm formation. *Coin* proposes a handle

Figure 5: Rule-governed uptake of emergent terms in Werewolf. The keynote traces a contested early-day frame (`baseline day`), a public reference pact for unresolved suspects (`pressure pool`), a trust-alignment rule (`core lock`), and a role-claim influence warning (`claim gravity`).

for local structure and tests it through same-turn, use-bound definition. *Share* places the accepted entry into a shared or visibility-restricted commons. *Use* tests whether another agent can reuse, refine, contest, or redeploy it under the local criterion of correct use. In grounding-theoretic terms, communication is not complete when a message is produced; it requires mutual evidence relative to the participants' joint purpose (Clark and Brennan, 1991). COGENT makes this process observable in multi-agent reasoning.

This lifecycle clarifies why metrics must be interpreted together. Lower post-warm-up token count suggests shorter later interaction, but not grounding by itself. Higher accuracy suggests useful structure preservation, but not the linguistic mechanism. High reuse indicates uptake, but only grounded use preserves the relevant task rule. A term introduced once and never reused is a label without uptake; a term reused incompatibly is surface alignment without semantic alignment; a term reused, challenged, repaired, and stabilized around a task-relevant rule becomes a public reasoning resource. This is the MAS design question posed by COGENT: agents may improve not only how much they exchange, but what their conversation is made of.

7 Conclusion

We introduced COGENT, a coin–share–use communication layer for LLM-based multi-agent systems. Agents coin task-local terms, bind them to definitions, use them immediately, share accepted entries through a shared term board, and test them in later interaction. The core claim is both linguistic and computational: a term becomes useful not through private invention alone, but through public uptake under shared task criteria.

Across MMLU, FOLIO, and GSM8K pilot slices, COGENT improves over vanilla debate on most model–task pairs and reduces post-warm-up debate tokens across all evaluated pairs, while showing that arithmetic tasks require coined terms to preserve exact verification steps. The Werewolf analysis extends the claim to hidden-information social reasoning: public terms form a compact, reusable, cross-speaker layer of convention under asymmetric common ground (Clark and Marshall, 1981). These findings suggest that LLM-MAS design is not only a matter of adding agents, rounds, or context. It also depends on whether agents can redesign the unit of interaction itself, turning repeated inferential structures into shared task-local words that others can use, contest, and stabilize.

Limitations

This work has several limitations. First, the main benchmark results use cost-controlled slices of MMLU, FOLIO, and GSM8K rather than full suites; small gains, especially on MMLU, should be treated as descriptive until paired significance tests, invalid-answer rates, and confidence intervals are reported. Second, COGENT changes multiple factors at once: coining passes, definition-and-use constraints, sharing, and lexicon-conditioned interaction. Stronger causal evidence requires call-matched debate, private-only coinage, public lexicon without forced reuse, and log-only coinage ablations. Third, our token column measures post-warm-up downstream debate output, not total API cost. COGENT may shorten later interaction while adding coining calls, prompt tokens, lexicon-injection overhead, latency, and monetary cost. Fourth, the Werewolf study is exploratory: game length and survival are path-dependent, and convention-formation comparisons require post-hoc pseudo-lexicons for the no-COGENT condition. Finally, lexical-function categories, novelty filters, and blocklists depend on design choices and should be validated with multi-annotator studies across more tasks, languages, and model families.

References

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC.

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Chidera Onochie Ibe, Srihas Rao, Arthur Caetano, and Misha Sra. 2026. Artificial leviathan: Exploring social evolution of LLM agents through the lens of hobbesian social contract theory. *Frontiers in Physics*, 14.

Ali Dorri, Salil S. Kanhere, and Raja Jurdak. 2018. Multi-agent systems: A survey. *IEEE Access*, 6:28573–28593.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 2137–2145. Curran Associates, Inc.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Simon Garrod and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8–11.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. 2024a. LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*. ArXiv:2402.03578v3.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024b.

FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, volume 30, pages 2146–2156. Curran Associates, Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (Natural) language. In *International Conference on Learning Representations*.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "Mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*, 1:9.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1495–1502.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–225.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3–4):379–423, 623–656. Parts I and II; Part II DOI: 10.1002/j.1538-7305.1948.tb00917.x.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Reid G. Smith. 1980. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, C-29(12):1104–1113.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford. Translated by G. E. M. Anscombe.

Michael Wooldridge. 2009. *An Introduction to MultiAgent Systems*, 2 edition. John Wiley & Sons, Chichester, UK.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *First Conference on Language Modeling*.

Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. 2025. Beyond self-talk: A communication-centric survey of LLM-based multi-agent systems. *arXiv preprint arXiv:2502.14321*. ArXiv:2502.14321v2.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. 2025. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*.

A Implementation Details

The main text treats COGENT as a coin–share–use communication layer around sequential debate. Table 3 expands the prompt abstractions used by the controller, and Algorithm 2 gives the vanilla sequential debate baseline. The baseline uses the same independent Round 0 and sequential debate order as COGENT, but it does not expose a lexicon channel.

Algorithm 2 Vanilla Sequential Debate$(q; R)$

Input: instance q, meta-rounds R
Output: final answer $\hat{a}$
1: initialize histories; $m_i^0 \leftarrow \text{LLMStep}(i, P_{\text{init}}(q))$ for $i \in \{A, B\}$
2: **for** $r = 1$ to R **do**
3: $\quad m_A^r \leftarrow \text{LLMStep}(A, P_{\text{orig}}(q, m_A^{r-1}, m_B^{r-1}))$
4: $\quad m_B^r \leftarrow \text{LLMStep}(B, P_{\text{orig}}(q, m_B^{r-1}, m_A^r))$
5: **end for**
6: $\hat{a} \leftarrow \text{Finalize}(m_A^R, m_B^R)$

B Metric Definitions

Final prediction is obtained from the latest two debate outputs, not from all answer markers in the transcript. For GSM8K, multiple-choice tasks, and FOLIO, the controller extracts one prediction from AgentA's latest debate response and one prediction from AgentB's latest debate response. If both predictions are valid and equal under the dataset-specific equality predicate, the shared answer is returned. Otherwise, the controller applies the dataset adapter's majority-vote rule over valid latest predictions only.

Let $\text{tok}(\cdot)$ denote the tokenizer used by the metric script. For a single-agent raw answer, if y_n^{raw} is the final output for instance n, the average token count is

$$\text{ATC}_{\text{raw}} = \frac{1}{|\mathcal{N}_{\text{raw}}|} \sum_{n \in \mathcal{N}_{\text{raw}}} \text{tok}(y_n^{\text{raw}}).$$

For vanilla debate, let $y_{i,n}^0$ be agent i's first real output before the first meta-round, and let $y_{i,n}^R$ be agent i's last real output in the final debate state. The baseline and final counts are

$$B_n^{\text{van}} = \sum_{i \in \{A,B\}} \text{tok}(y_{i,n}^0),$$
$$F_n^{\text{van}} = \sum_{i \in \{A,B\}} \text{tok}(y_{i,n}^R).$$

and $\text{CR}_n^{\text{van}} = F_n^{\text{van}} / B_n^{\text{van}}$ when logged. For COGENT, let $y_{i,n}^{\text{wu}}$ be the last real output of agent i before the debate marker inside the final coining-active meta-round. Then

$$B_n^{\text{cog}} = \sum_{i \in \{A,B\}} \text{tok}(y_{i,n}^0),$$
$$W_n^{\text{cog}} = \sum_{i \in \{A,B\}} \text{tok}(y_{i,n}^{\text{wu}}).$$

and $\text{CR}_n^{\text{cog}} = W_n^{\text{cog}} / B_n^{\text{cog}}$. The token column for COGENT is computed from the final no-compression meta-round after warm-up.

C Lexical-Function Annotation Protocol

For the soft lexical-function analysis in Figures 2 and 3, we use human annotation followed by LLM-assisted verification. We first constructed a six-category codebook from manual inspection of benchmark lexicons: discriminative cue, domain concept, quantity/unit variable, operation checkpoint, proof-status logic, and gap/uncertainty marker. A human annotator then assigned soft functional labels to coined problem terms, allowing a term to receive fractional weight across multiple categories when its use overlapped functions. An LLM verifier was subsequently given the term, definition, local task context, and assigned category weights and asked to flag inconsistent or unsupported labels. Flagged cases were manually adjudicated; the LLM verification step was used as a consistency check rather than as the final labeling authority. Percentages are normalized over problem-term counts within each task and model.

D Lexical Ecology Details

Table 4 reports the detailed task-by-task lexical ecology summarized in Section 6.2. It preserves the full lexicon scales, criteria of correct use, and illustrative terms for each benchmark and for the exploratory Werewolf setting.

E Additional Werewolf Lexical-Grounding Metrics

Table 2 in the main text reports the compact Werewolf signals. Tables 5, 6, and 7 retain the auxiliary function, comparison, footprint, productivity, and persistence diagnostics.

F Generative AI Statement

Generative AI tools were used in this work only to support non-substantive writing and presentation

Symbol	Function
P_{init}	Dataset-specific Round-0 solving prompt. In the default setting it does not include the lexicon field, so the initial answer is treated as ordinary independent reasoning rather than as a coining turn.
P_{judge}	Private coining header containing the previous response statistics, the agent id, visible terms, cumulative lexical counts, and the previous-turn word count.
P_{name}	Private coining prompt. It asks for `[PART 1]` compact reasoning and `[PART 2]` new named terms, forbids `[PART 3]` and final-answer markers, and requires every term defined in `[PART 2]` to be used in `[PART 1]`.
P_{debate}	Public use prompt conditioned on the opponent answer and the detailed `KNOWN_TERMS` block. If public terms exist, the prompt requires at least one of them to be considered or reused. AgentA sees the opponent's pre-debate answer, while AgentB sees AgentA's fresh answer from the current debate round.

Table 3: Full prompt abstractions in COGENT.

Task	Lexicon scale and dominant function	Criterion of correct use	Illustrative terms
MMLU	3411 entries; 2759 unique terms; classificatory pacts for option-level contrasts.	The term must preserve the distinction that makes one answer preferable to its strongest distractors.	expansiovision; treaty-fulfill mandate; focus-exclude rule; cue-power fading
GSM8K	3128 entries; 1895 unique terms; task-local notation for quantities, roles, and operations.	The term must preserve units, variable bindings, operation structure, and checkable intermediate steps.	commissionalize; speedtime; unitcost; pumpcost; josh cookies
FOLIO	644 entries; 461 unique terms; proof-status operators for rule-governed inference.	The term must preserve whether the conclusion is forced, contradicted, or left open by the premises.	open-model gap; must-follow path; contradiction path; linkvoid; lacksupport
Werewolf	32 accepted entries; 417 reuses; social-epistemic grounding under hidden information.	The term must preserve trust, suspicion, role-claim evidence, voting accountability, and visibility boundaries.	baseline day; pressure pool; core lock; claim gravity

Table 4: Lexical ecology across the answer-oriented benchmarks and the exploratory Werewolf language game. The relevant unit is not merely the number of coined strings, but the task-specific criterion under which a term can be used correctly.

tasks, including formatting, generating LaTeX templates, checking grammar, improving clarity, and refining word choice. These tools were not used to replace the authors' scientific judgment, conduct independent analysis, or determine the study's claims. The authors carefully reviewed, edited, and verified all AI-assisted content to ensure factual accuracy, consistency with the reported experiments, and academic integrity.

Function	Count	Criterion of correct use
Hidden-information management	9	Preserve visibility boundaries around night actions, protection, potions, and hidden roles.
Role-claim evidence	7	Track claims, counterclaims, checks, and the evidential force of role speech acts.
Deception strategy	6	Mark wolf cover, blending, safe echoing, or misdirection rather than ordinary disagreement.
Process structure	4	Organize day discussion, pressure allocation, and unresolved scrutiny.
Voting action	3	Connect public reasoning to banishment choices and vote accountability.
Trust alignment	2	Stabilize soft-clears, core trust, and late-game coalition maintenance.
Other social reasoning	1	Capture residual social-inference patterns not covered above.

Table 5: Functional ecology of accepted Werewolf terms.

Comparison axis	ORIGIN/no-COGENT	COGENT	Interpretation
Process-structure density	7.07 / 1k tokens	9.42 / 1k tokens	COGENT more often names and reuses structures for organizing discussion.
Accountability-revision density	5.33 / 1k tokens	6.24 / 1k tokens	COGENT more often revises responsibility after votes or claims.
Common-prefix public-log tokens	37,593	35,628	Matched-checkpoint public context decreases by 5.23%.
Common-prefix all-player personal-log tokens	231,330	219,873	Matched-checkpoint personal context decreases by 4.95%.
Full-game total tokens	763,217	1,288,361	Not an efficiency win; game length and additive lexicon injection increase total cost.
Explicit definition-bound terms	Post-hoc only	32 accepted entries	COGENT makes convention formation auditable rather than implicit.
Observed known-term reuses	Not applicable by design	417	The explicit lexicon channel makes uptake measurable.
Cross-speaker known-term uptake	Requires pseudo-lexicon baseline	15 / 32 terms used by ≥ 2 public speakers	Evidence of public uptake, with a fair direct baseline left to post-hoc convention extraction.

Table 6: Werewolf comparison with ORIGIN/no-COGENT.

Metric	Value	Interpretation
Accepted Werewolf terms	32	A small task-local lexicon is formed across five hidden-role games.
Observed term reuses	417	Accepted terms are taken up repeatedly in later interaction rather than remaining one-off labels.
Lexical uptake rate	13.03	Each accepted term is reused 13.03 times on average.
Final lexicon snapshot tokens	771	The stored vocabulary remains compact.
Lexicon footprint ratio	0.031	Final lexicon snapshots occupy about 3.1% of the final COGENT public-log footprint.
Reuse per 100 lexicon tokens	54.09	The lexical layer is productive relative to its token footprint.
Public terms / reuse	12 / 290	Public entries are fewer than wolf-private entries but generate most observed reuse.
Wolf-private terms / reuse	20 / 127	Faction-private terms are more numerous but less productive per entry.
Public reuse share	69.5%	Public terms contribute most observed reuse despite being 37.5% of entries.
Public uptake advantage	$3.81\times$	Reuse per public entry is about 3.8 times reuse per wolf-private entry.
Terms used in public speech	22 / 32	Most accepted terms appear in public discussion.
Terms used by ≥ 2 public speakers	15 / 32	Nearly half of accepted terms show cross-speaker public uptake.
Avg. unique public speakers per term	2.375	Term use is not confined to a single speaker.
Public checkpoint persistence	0.472	A substantial fraction of public terms persists across later checkpoints.

Table 7: Full exploratory Werewolf lexical-grounding metrics.