SELF-COINED TERMS IMPROVE COMMUNICATION IN MULTI-AGENT LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While LLM-based multi-agent systems (MASs) have emerged as a promising paradigm for solving complex tasks, agents often restate the same concepts using different expressions across rounds, leading to inter-agent misunderstanding and redundant communication. Inspired by how humans create new words to facilitate communication, we hypothesize that LLM agents possess a similar ability. To address the challenge, we introduce COGENT (**Co**llective **G**rounding of **E**mergent **N**amed **T**erms), a training-free, effective, and efficient communication layer in which agents communicate using self-coined terms. Each agent first coins task-specific terms together with their definitions and uses these terms in its own reasoning. The coined terms are then posted to a shared term board and reused in subsequent debate rounds. This design aligns agents' representations of recurring concepts and reduces redundant communication. Extensive experiments with five models, including GPT, Qwen, Claude, and Gemini, show that COGENT consistently improves accuracy over vanilla debate across diverse dataset-model combinations on six benchmarks: MMLU, FOLIO, GSM8K, MedQA, GPQA-Diamond, and HLE, while also reducing debate tokens after the term-coining stage. On HLE, COGENT improves Claude-4.6-Sonnet's accuracy from 18.33% to 27.00% while reducing final-round debate tokens by 56.5% relative to vanilla debate. A Werewolf test and a visibility ablation further demonstrate that the full public-sharing protocol leads to denser reuse of coined terms than vanilla communication. Our code and data will be released upon acceptance.

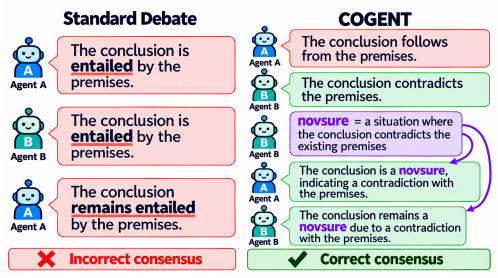


Figure 1: A motivated example. Agents coin and use a new term for their debate.

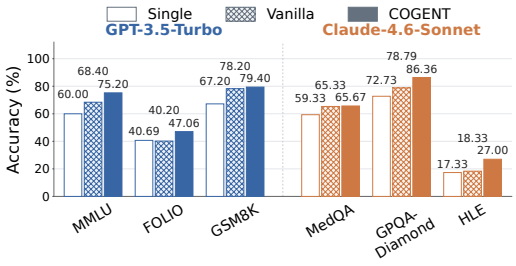


Figure 2: COGENT improves vanilla debates in various baselines.

1 INTRODUCTION

In LLM-based multi-agent systems (MASs), a key bottleneck is 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 communicate by passing free-form natural-language traces to critique answers or revise decisions (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). Although LLMs can coin new expressions on their own, these protocols give such expressions no explicit definition, no shared registry, and no expectation of reuse. Such traces are expressive, but weak as reusable infrastructure: the same proof status or option contrast must be rephrased across turns, which can lengthen interaction or drift in meaning. Figure 1 illustrates this on FOLIO: vanilla debate converges

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

on the wrong TRUE, whereas under COGENT a shared term (`novsure`) lets the agents reach the correct FALSE.

This paper studies lexical innovation as a communication primitive for 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 & Marshall, 1981; Lewis, 1969; Brennan & Clark, 1996). Finally, a local expression becomes globally useful when others can apply it under the same task scenarios. We hypothesize that LLMs have also learned this capability from human language and that exploiting it can make multi-agent communication more efficient. We therefore ask *whether LLM agents can similarly coin task-local terms, expose them to collaborators, and use them as shared reasoning units.* Figure 2 previews the payoff: on every displayed model–benchmark pair, debate with shared self-coined terms is more accurate than vanilla debate.

We introduce COGENT, **Co**llective **G**rounding of **E**mergent **N**amed **T**erms, a communication layer organized around three steps: *Self-coin*, *Share*, and *Use* (Figure 3). In Self-coin, each agent creates a short task-specific term in the form `term = brief definition` and uses it in the same response. In Share, accepted terms are added to the shared term board, which agents see as a `KNOWN_TERMS` block. In Use, agents debate by applying the term as written, questioning it, or refining it. Following use-based accounts of meaning, a term matters because agents can use it consistently within the task, not simply because it is short (Wittgenstein, 1953; Grice, 1975; Clark & Marshall, 1981; Clark & Brennan, 1991; Brennan & Clark, 1996; Lewis, 1969; Pickering & Garrod, 2004).

We evaluate COGENT on MMLU, FOLIO, GSM8K, MedQA, GPQA-Diamond, and HLE with GPT-3.5-Turbo, Qwen3.5-35B-A3B, Claude-4.6-Sonnet, and Gemini-3.5-Flash, and further test it in a Werewolf game simulation. COGENT improves accuracy over vanilla debate on average and reduces final-round debate tokens. On HLE, COGENT improves Claude-4.6-Sonnet's accuracy from 18.33% to 27.00% while reducing final-round debate tokens by 56.5% relative to vanilla debate. A component ablation shows that removing definitions, cross-agent sharing, or enforced reuse each lowers accuracy, and a Werewolf visibility ablation shows that sharing, rather than term volume, drives reuse: a faction-private variant accepts almost as many terms (109 vs. 114), yet observed reuse falls from 776 to 337. Overall, we shift MAS design from exchanging more traces to creating shared task-local terms that make later communication more stable and efficient.

Our contributions are three-fold:

- We introduce COGENT, a new communication paradigm for LLM-based multi-agent systems.
- Across six reasoning benchmarks and multiple frontier and open-source LLMs, we show that COGENT consistently improves multi-agent reasoning accuracy over vanilla debate while substantially reducing communication length.
- Through game-based evaluation and controlled ablations, we analyze how emergent terms are created, adopted, and reused across agents.

2 RELATED WORK

Communication protocols in LLM-based multi-agent systems. Classical MAS research treats 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 MASs increasingly treat communication as interaction rather than a single forward pass, such as multi-agent debate (Du et al., 2024) and multi-model discussion with confidence-weighted voting (Chen et al., 2024). A parallel line of work formalizes the transport layer itself: interoperability protocols such as MCP, ACP, A2A, and ANP standardize how agents discover one another and exchange messages (Ehtesham et al., 2025), and a technical taxonomy of nine open-source protocols finds the design space converging on session-state persistence, hybrid payloads, and flexible schemas rather than on a single standard (Sander et al., 2026). 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 efforts standardize the *envelope* of

agent communication—transport, discovery, session state—while leaving the *content* of a message an unstructured natural-language string. CoGENT is complementary: it asks what reusable objects should travel inside that envelope.

Grounding, conceptual pacts, and lexical coordination. CoGENT connects MAS design to theories of dialogue grounding and lexical coordination. Work on semantic coordination, conceptual pacts, and interactive alignment shows that interlocutors establish local conventions shaped by the current joint activity (Grice, 1975; Lewis, 1969; Clark & Marshall, 1981; Garrod & Anderson, 1987; Clark & Brennan, 1991; Brennan & Clark, 1996; Pickering & Garrod, 2004; Garrod & Pickering, 2004): a local name becomes efficient only after participants know how to use it. Emergent-communication work studies learned symbol systems under multi-agent rewards (Foerster et al., 2016; Lazaridou et al., 2017; Havrylov & Titov, 2017; Mordatch & Abbeel, 2018; Stengel-Eskin et al., 2026) extend this to LLM populations, showing that language does evolve between LLM agents under efficiency pressure but drifts toward codes their human readers can no longer interpret. CoGENT takes the opposite design decision: every accepted entry carries an explicit `term = definition` pair and must be used in the same response, so the emergent lexicon stays auditable by collaborators and by us. Closest to our setting, Pei et al. (2026) let agents evolve compact symbolic protocols offline and route them at inference time, whereas CoGENT coins terms per instance inside the ongoing dialogue and treats definition-bound public uptake as the unit of analysis. Finally, Sun et al. (2025) treat clarification questions as entropy reduction over an under-specified instruction; CoGENT applies the same logic inside a peer group rather than between user and agent, since a visible term is tested by being questioned or refined, not merely repeated. In this sense, CoGENT re-frames multi-agent communication from exchanging more text to building shared words that structure later reasoning.

Compression, memory, and reusable reasoning state. Long deliberation between agents increases context length and makes later turns depend on verbose 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.

3 Method: The Self-coin–Share–Use Communication Layer

Ordinary debate recreates the same distinctions in new words each turn; CoGENT instead lets agents name a distinction once, expose the name and its definition to a partner, and hold both accountable for using it consistently afterward. CoGENT is a communication layer around a base LLM f_θ (Vaswani et al., 2017). For each problem q, it creates two agents, Agent A and Agent B, with separate histories, private self-coined lexicons L_A and L_B, and a shared term board L^{pub}. Both agents first answer independently, matching the starting point of vanilla debate. They then run R meta-rounds, each with three steps: Self-coin creates task-specific terms, Share places accepted terms on the shared term board, and Use brings those terms into later debate. Self-coin is active only in the first R_c meta-rounds, which we call the *warm-up*; the remaining $R - R_c$ meta-rounds are debate-only (*post-warm-up*). Figure 3 gives an overview, Algorithm 1 specifies the details, and Table 8 summarizes the prompts. $\text{LLMStep}(i, \cdot)$ calls f_θ on agent i's own history, which we omit from the equations below.

Pre-step: ordinary independent solving. Each agent first answers independently, $m_i^0 = f_\theta(P_{\text{init}}(q))$ for $i \in \{A, B\}$, where P_{init} is the dataset-specific Round-0 solving prompt without any lexicon field (Table 8, Appendix A). CoGENT and vanilla debate therefore begin from the same ordinary answers.

Step 1: Self-coin. During the first R_c meta-rounds, each agent performs K coining passes ($k = 1, \ldots, K$) over its own latest response:

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

Algorithm 1 CoGENT communication

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

$L_A, L_B, L^{\text{pub}} \leftarrow \emptyset$
Initialize agent histories
Pre-step: independent solving
For each $i \in \{A, B\}$:
$m_i^0 \leftarrow \text{LLMSTEP}(i, P_{\text{init}}(q))$
$\ell_A, \ell_B \leftarrow m_A^0, m_B^0$
for $r = 1$ to R **do**
$K_r \leftarrow K$ if $r \le R_c$, otherwise 0
Self-coin
for all $i \in \{A, B\}$ **do**
$(L_i, \ell_i) \leftarrow \text{SELFCOIN}(i, \ell_i, L_i, L^{\text{pub}}, K_r)$
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
$\hat{a} \leftarrow \text{FINALIZE}(m_A^R, m_B^R)$

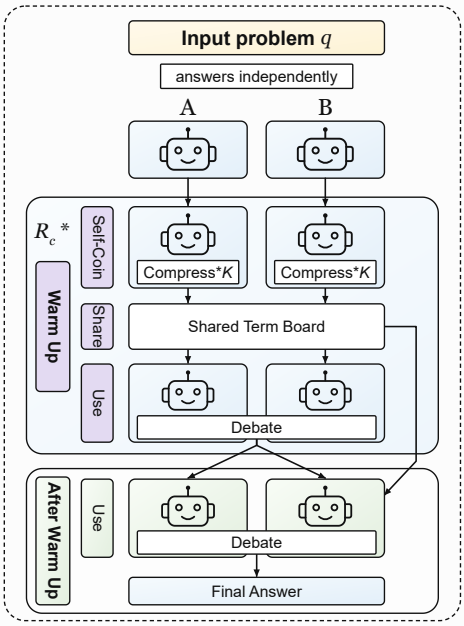


Figure 3: Workflow of CoGENT. Self-coin is active during the first R_c rounds; subsequent rounds continue with shared-term-guided debate.

Here $\ell_i^{r,k-1}$ is agent i's latest response ($\ell_i^{r,0}$ is its response entering meta-round r), $\ell_i^{r,k}$ is the rewritten response returned by pass k, $L_i^{r,k-1}$ is its private lexicon so far, L_{r-1}^{pub} is the shared term board from the previous meta-round, and $\Delta L_i^{r,k}$ is the set of terms newly accepted in pass k, so that $L_i^{r,k} = L_i^{r,k-1} \cup \Delta L_i^{r,k}$. P_{name} is the private coining prompt, prefixed by a header P_{judge} that reports the agent id, visible terms, cumulative lexical counts, and the previous-turn word count (Table 8). During Self-coin, each agent reviews its latest reasoning and builds a private list of task-specific terms. A candidate is accepted only if it has the form $t = d$ (a term t with a brief definition d), is used in the same response, passes the novelty and blocklist checks, and contains no final-answer marker. Same-response use shows that the term already plays a role in the creator's reasoning. This differs from ordinary summarization, which can shorten a response without creating a named unit for later discussion.

Step 2: Share. After Self-coin in each meta-round, accepted entries from the agent-private lexicons are shared through the shared term board:

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

where L_i^r is agent i's lexicon after its last coining pass; each agent's post-Self-coin response $p_i^r = \ell_i^{r,K_r}$ ($K_r = K$ during warm-up and 0 afterwards) is the message its partner sees in Use. Each shared entry records the term, its definition, and who created it. Share makes the entry visible to both agents, but visibility alone does not show that the entry is understood or used. That evidence comes from later interaction.

Step 3: Use. Use is a sequential debate guided by the shared term board. When the board is nonempty, the prompt asks each debate response to explicitly reuse at least one visible term. The agent may accept, question, or refine the term it uses; other visible entries may be ignored. The sequential updates are

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

where m_i^r is agent i's public debate message in meta-round r and P_{debate} is the public Use prompt (Table 8). Agent A responds to p_B^r, whereas Agent B responds to Agent A's new message m_A^r; after each Use turn, the speaker's lexicon and latest response are updated and the board is refreshed (Algorithm 1). A visible term that is never reused remains untested. Later reuse, questioning, or refinement provides evidence that the term has entered the public interaction; cross-speaker reuse provides stronger evidence that it has become shared.

Post-step: finalization. The controller parses the answer lines of the two latest Use messages and returns $\hat{a} = \textsc{Finalize}(\text{Extract}(m_A^R), \text{Extract}(m_B^R))$, applying the agreement/vote rule in Appendix B. Finalization only reads predictions and is kept separate from the Self-coin–Share–Use process.

Metrics. We report task accuracy, compression rate, and post-warm-up token count. Let $\text{tok}(\cdot)$ be the tokenizer used throughout (Appendix B). For CoGENT, let B be the two agents' summed initial-output token count and W their summed token count on their last warm-up outputs, i.e., $\text{tok}(\cdot)$ applied to each agent's last output before the debate-only rounds; the compression rate is then $\text{CR} = W/B$. For VANILLA, W is replaced by the final-round output count (Appendix B). The Tok. column instead reports output tokens averaged over the final, debate-only (post-warm-up) meta-round, so it differs from W; it is not total API usage. For Werewolf, where there is no single gold answer, we additionally report lexical uptake $\text{Uptake} = N_{\text{reuse}}/N_{\text{term}}$, where N_{reuse} is the number of observed term reuses and N_{term} is the number of accepted terms. For visibility comparisons, where Werewolf terms sit in a public pool (pub) or a wolf-private pool (wolf; Section 5), we also report the public reuse share $N_{\text{reuse}}^{\text{pub}}/N_{\text{reuse}}$ and the public uptake advantage $\text{PublicAdv} = (N_{\text{reuse}}^{\text{pub}}/N_{\text{term}}^{\text{pub}})/(N_{\text{reuse}}^{\text{wolf}}/N_{\text{term}}^{\text{wolf}})$; values above one indicate that public entries are reused more often per entry than wolf-private entries.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate CoGENT on six answer-oriented benchmarks: MMLU (Hendrycks et al., 2021), FOLIO (Han et al., 2024b), GSM8K (Cobbe et al., 2021), MedQA (Jin et al., 2021), GPQA-Diamond (Rein et al., 2023), and HLE (Phan et al., 2025), using cost-controlled slices of 500, 204, 500, 300, 198, and 300 examples. Backbones are GPT-3.5-Turbo and Qwen3.5-35B-A3B with thinking disabled for the original block, and Claude-4.6-Sonnet and Gemini-3.5-Flash under 4096- and 8192-token generation budgets for the added block. Each model–benchmark pair compares SINGLE, a one-agent answer; VANILLA, two-agent sequential debate without a lexicon channel; and CoGENT, which adds the Self-coin–Share–Use layer with $(R, K, R_c) = (3, 2, 2)$. We additionally report a component-focused MMLU ablation with six intermediate conditions between the VANILLA and Full anchors, a cross-framework check that layers CoGENT onto Multiagent Debate and RECONCILE with Kimi-K2.6 (Du et al., 2024; Chen et al., 2024), and a five-game Werewolf stress test with a 50-seed visibility ablation. Appendices A–G give the full designs, prompt summaries, and annotation protocols.

4.2 RESULTS

Table 1 compares CoGENT with vanilla debate. CoGENT improves accuracy on five of six model–task pairs; the two largest gains, MMLU and FOLIO on GPT-3.5-Turbo ($+6.80$ and $+6.86$ points), are consistent with coined terms naming domain concepts on the former and marking proof status on the latter. The only regression is Qwen3.5-35B-A3B on GSM8K ($96.59\% \rightarrow 96.00\%$), where stronger shortening coexists with a small loss in exact arithmetic accuracy and marks a boundary of the mechanism. Post-warm-up debate tokens fall in every pair, by 12.9–70.4% for GPT-3.5-Turbo and 35.1–74.3% for Qwen3.5-35B-A3B.

Table 2 extends the evidence to stronger models and more heterogeneous datasets. CoGENT improves over VANILLA on all six added pairs: $+0.34$, $+7.57$, and $+8.67$ points for Claude-4.6-Sonnet on MedQA, GPQA-Diamond, and HLE, and $+0.67$, $+3.53$, and $+6.67$ points for Gemini-3.5-Flash, while cutting final debate tokens by 65.4%, 86.7%, and 56.5% and by 52.9%, 42.5%, and 34.8%

Table 1: Results on MMLU, FOLIO, and GSM8K. ↑/↓: higher/lower is better. *CR*: compression rate. *Tok.*: output tokens for SINGLE and post-warm-up debate tokens for VANILLA/COGENT.

Model	Variant	MMLU			FOLIO			GSM8K		
		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	68.40	0.95	361.84	40.20	0.99	195.74	78.20	1.97	99.22
	COGENT	**75.20**	**0.37**	**107.29**	**47.06**	**0.46**	120.74	**79.40**	**0.79**	**86.47**
Qwen3.5-35B-A3B w/o thinking	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 2: Results on MedQA, GPQA-Diamond, and HLE. Columns as in Table 1.

Model	Variant	MedQA			GPQA-Diamond			HLE		
		Acc.↑	CR↓	Tok.↓	Acc.↑	CR↓	Tok.↓	Acc.↑	CR↓	Tok.↓
Claude-4.6-Sonnet	Single	59.33	N/A	216.07	72.73	N/A	828.87	17.33	N/A	951.78
	Vanilla	65.33	0.92	389.00	78.79	0.66	895.11	18.33	0.65	988.76
	COGENT	**65.67**	**0.31**	**134.42**	**86.36**	**0.24**	**119.22**	**27.00**	**0.09**	**430.39**
Gemini-3.5-Flash	Single	94.33	N/A	425.80	82.83	N/A	321.85	26.33	N/A	**240.00**
	Vanilla	95.33	1.11	891.45	86.87	2.41	1022.13	38.33	1.18	584.37
	COGENT	**96.00**	**0.20**	**419.51**	**90.40**	**0.39**	**587.34**	**45.00**	**0.42**	381.03

respectively. Averaged over the six added pairs, accuracy rises from 63.83% to 68.41% and average final debate tokens drop from 795.14 to 345.32, a 56.6% reduction.

4.3 CROSS-FRAMEWORK GENERALIZATION

We test whether the lexical layer is tied to our own sequential debate wrapper by applying it to two MMLU communication backbones with moonshotai/Kimi-K2.6 (Table 3). Adding COGENT to the multiagent debate protocol of Du et al. (2024) leaves accuracy at the single-agent level while cutting final debate tokens by about three quarters; adding it to a RECONCILE-style round-table consensus protocol (Chen et al., 2024) recovers 8.60 accuracy points at a comparable trace length. The two backbones therefore benefit along different axes: compression where debate is already accurate, accuracy where consensus is weak.

Table 3: Cross-framework results on MMLU with Kimi-K2.6. Different prompts and aggregation rules make this a portability check, not a framework ranking.

Setting	Acc.↑	CR↓	Tok.↓
SINGLE	93.00	N/A	489.02
MAD	93.00	1.23	659.87
+ COGENT	**93.60**	**0.26**	**172.20**
RECONCILE	79.60	0.84	408.82
+ COGENT	**88.20**	**0.82**	**402.46**

Table 4: Lexicon scale across six benchmarks. All entries are shared. Unique terms are counted within each task and summed in Total. Werewolf is excluded.

Task	Entries	Unique	Uniq./Ent.
MMLU	3,411	2,759	80.9%
GSM8K	3,128	1,895	60.6%
FOLIO	644	461	71.6%
MedQA	6,443	5,952	92.4%
GPQA-Diamond	4,005	3,619	90.4%
HLE	7,237	6,830	94.4%
Total	**24,868**	**21,516**	**86.5%**

Table 5: Representative coined terms with their logged definitions.

Task	Representative logged term–definition entries
MMLU	`expansiovision` = "belief in US expansion destiny"; `treaty-fulfill mandate` = "principle enforcing treaty obligations"; `focus-exclude rule` = "concentrating on one stimulus while disregarding others"
GSM8K	`commissionalize` = "Calculate the commission amount from a sale"; `speedtime` = "rate multiplied by time"; `josh cookies` = "number of cookies Josh made"
FOLIO	`open-model gap` = "a logical state where both a statement and its negation are consistent with premises"; `must-follow path` = "conclusion is entailed by all models"; `contradiction path` = "a logical trajectory where conclusion is falsified"
MedQA	`xynorps` = "the presence of active delusions or hallucinations for at least 2 weeks in the absence of a major mood episode"; `helicotriple` = "H. pylori triple therapy comprising proton pump inhibitor and two antibiotics"; `parathyrtilt` = "parathyroid involvement acting as the pivotal differentiator between MEN 2A and 2B"
GPQA-Diamond	`krillax` = "the minimum energy separation required to distinguish two quantum states, defined by the sum of their individual line widths"; `ylidcyc` = "sulfoxonium ylide reaction converting synket to cyclopropanated synprop (adding one carbon)"; `combival` = "the combined expectation value of the operator $10\sigma_z + 5\sigma_x$, equal to approximately -0.67"
HLE	`exactmatch-flag` = "indicator that the answer must match a precise string/entity"; `dbgap` = "inability to identify the exact target paper/answer without database access"; `zorbkey` = "orbital setup shorthand: given periapsis/apoapsis altitudes $\rightarrow r_p, r_a, a, e$"

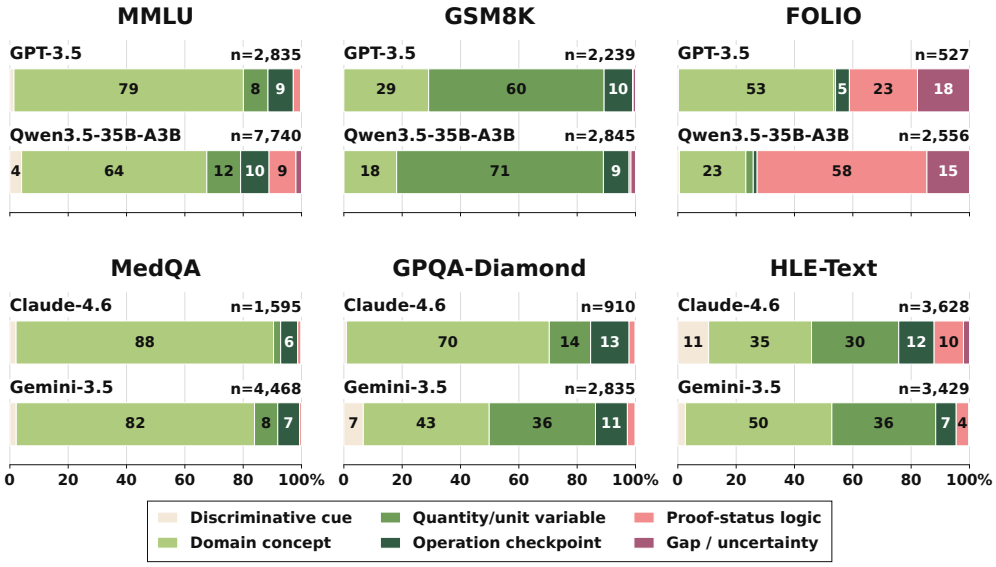


Figure 4: Lexical functions of coined terms by task and model. Bars give the percentage of problem–term counts in each of six categories; n is the problem–term count (codebook in Appendix E).

4.4 What the coined terms do

Table 4 reports accepted entries, de-duplicated surface forms, and shared-board exposure for the six answer-oriented benchmarks, summing to 24,868 accepted entries and 21,516 task-level unique terms. Table 5 then audits a small set of representative coined terms per task, showcasing coined terms of diverse usages (Appendix F lists the full per-task discussion). Figure 4 shows how the terms specialize by task (codebook in Appendix E): domain concepts lead on MMLU, MedQA, and GPQA-Diamond, quantity variables dominate GSM8K, and proof-status and gap markers concentrate on FOLIO.

Figure 5 contrasts independent VANILLA and COGENT runs on two harder benchmarks (the FOLIO case appears in Figure 1). In GPQA-Diamond, the agents split B/C in Round 0; Agent A coins `zelvorn` for anti-Hermiticity, and after *Share* Agent B uses it to recheck the disputed matrix entries and switches from C to B, whereas both VANILLA agents keep the misread entry and converge on C. In HLE, Agent B coins `evengridform` for the exact even-n grid formula, and Agent A applies it at $n = 14$ to replace the candidate 75 with 76; the VANILLA run repeats 75 and ends without a valid

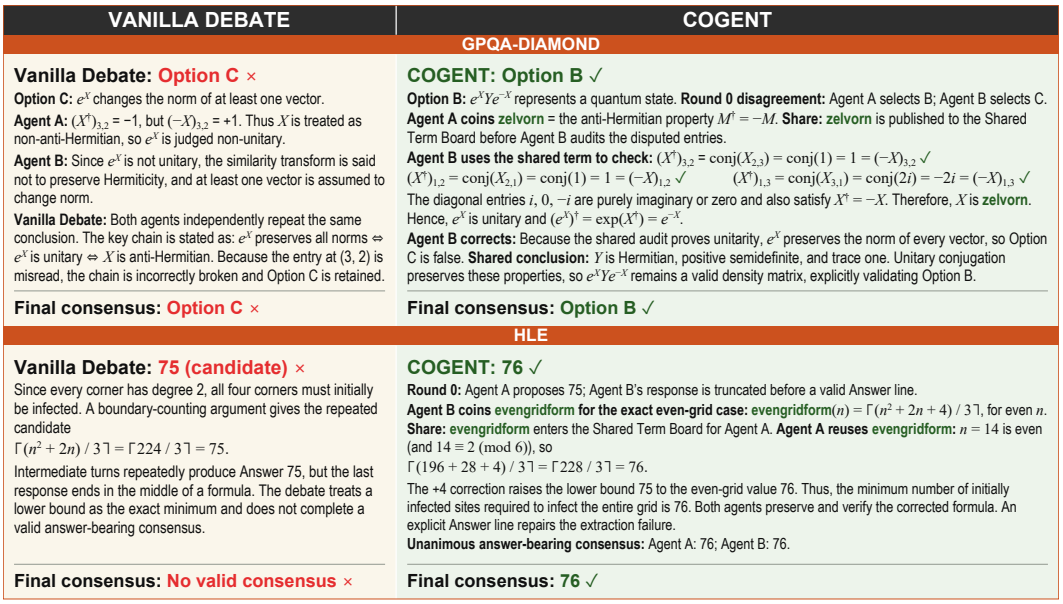


Figure 5: Two logged cases on GPQA-Diamond (top) and HLE (bottom), comparing independent VANILLA and COGENT runs; full panels with problem statements are in Figure 8 (Appendix C).

Table 6: Rule-governed uptake of four emergent terms in Werewolf.

Metric	Value
Accepted terms	32
Observed reuses	417
Terms used by	
≥ 2 public speakers	15 / 32
Public reuse share	69.5%
Prefix reduction: public	5.23%
Prefix reduction: personal	4.95%

Table 7: Visibility ablation on 50 GPT-4o-mini games (seeds 42–91). Full: COGENT; Priv.: faction-private ablation; Ratio: Full/Priv.

Metric	Full	Priv.	Ratio
Accepted terms	114	109	1.05×
Observed reuses	776	337	2.30×
Lexical uptake	6.81	3.09	2.20×
Reuse / 100 lex. tok.	38.26	18.36	2.08×
Reuse / 100k tok.	8.956	4.470	2.00×
Reuse / model call	0.1900	0.0917	2.07×
Public terms / reuses	54/470	33/94	–
Wolf-priv. / reuses	60/306	76/243	–
Public reuse share	60.57%	27.89%	2.17×
Public uptake adv.	1.71×	0.89×	1.92×

answer. Appendix C gives all three cases with problem statements and gold-answer checks. Each sequence shows cross-speaker use before stable agreement, not the causal effect of a single term.

5 WEREWOLF: CONVENTION FORMATION UNDER HIDDEN INFORMATION

5.1 SETUP AND RESULTS

Werewolf extends COGENT from answer-oriented reasoning to social-epistemic coordination under hidden information (Park et al., 2023; Dai et al., 2026). The criterion is interactional rather than propositional: a term must preserve provisional trust, role-claim evidence, unresolved pressure sets, and voting responsibility, and it is meaningful when later agents know when to apply it, when to reject it, and what social move it licenses. COGENT maintains a public term pool visible to all players and a wolf-private pool visible only to the wolf faction, so the same protocol can be observed under symmetric and asymmetric visibility.



Figure 6: 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`).

Table 6 summarizes the original five-game stress test. CoGENT accepts 32 terms and observes 417 reuses, a lexical uptake rate of 13.03. Public exposure is the productive part: public terms are 37.5% of accepted entries but account for 69.5% of observed reuse, a reuse-per-entry rate 3.81 times that of wolf-private terms. These numbers are not a clean full-game efficiency win; they show that CoGENT makes convention formation measurable under hidden information. A separate 50-seed GPT-4o-mini visibility ablation serves as a mechanism check (Table 7): the faction-private condition accepts nearly as many terms as full CoGENT (109 vs. 114), yet observed reuse falls from 776 to 337, and the gap persists per trajectory token and per model call (2.00× and 2.07×; Appendix G.1).

5.2 Rule-governed uptake

Figure 6 shows four such terms in action: `baseline day` becomes a contested early-day frame, `pressure pool` a shared reference pact for unresolved suspects, `core lock` a late-game trust-alignment rule, and `claim gravity` names how a role claim pulls 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, although the trace alone does not prove semantic understanding.

6 Conclusion

We introduced CoGENT, a communication layer organized around three steps: Self-coin, Share, and Use. The lexical analysis shows that self-coined terms take different forms across medical, scientific, logical, mathematical, and open-domain tasks. The Werewolf experiments further show that the full public-sharing protocol produces denser reuse than a similarly sized faction-private vocabulary. Together, these findings suggest that multi-agent systems can improve not only by exchanging more text, but also by creating a shared vocabulary that agents can use, question, and refine.

AI USE STATEMENT

Generative AI tools assisted with drafting and editing text and proofs, refining mathematical statements, literature discovery, and checking derivations and numerical identities. The preparation included separate AI-assisted reviews of proofs, cited sources, and presentation. However, the paper's final review of the content remains to be completed manually by the authors.

REPRODUCIBILITY STATEMENT

We will release all code, data, and traces upon acceptance. We have already listed the prompt, parameters, data examples, and settings in the appendix for the purpose of reproduction. Furthermore, we construct the code in a flexible way so that we can rapidly adapt to newer data or models.

REFERENCES

Fabrício Benevides, Jean-Claude Bermond, Hicham Lesfari, and Nicolas Nisse. Minimum lethal sets in grids and tori under 3-neighbour bootstrap percolation. *European Journal of Combinatorics*, 119:103801, 2024. doi: 10.1016/j.ejc.2023.103801. URL https://doi.org/10.1016/j.ejc.2023.103801.

Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996. doi: 10.1037/0278-7393.22.6.1482. URL https://doi.org/10.1037/0278-7393.22.6.1482.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, Washington, DC, 1991. doi: 10.1037/10096-006. URL https://doi.org/10.1037/10096-006.

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

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Chidera Onochie Ibe, Srihas Rao, Arthur Caetano, and Misha Sra. Artificial leviathan: Exploring social evolution of LLM agents through the lens of hobbesian social contract theory. *Frontiers in Physics*, 14:1700712, 2026. doi: 10.3389/fphy.2026.1700712. URL https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2026.1700712/abstract.

Ali Dorri, Salil S. Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *IEEE Access*, 6: 28573–28593, 2018. doi: 10.1109/ACCESS.2018.2831228. URL https://doi.org/10.1109/ACCESS.2018.2831228.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 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*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP). *arXiv preprint arXiv:2505.02279*, 2025. doi: 10.48550/arXiv.2505.02279. URL https://arxiv.org/abs/2505.02279.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2137–2145. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c7635bfd99248a2cdef8249ef7bfbef4-Abstract.html.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259, 2024. doi: 10.1057/s41599-024-03611-3. URL https://www.nature.com/articles/s41599-024-03611-3.

Simon Garrod and Anthony Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987. doi: 10.1016/0010-0277(87)90018-7. URL https://doi.org/10.1016/0010-0277(87)90018-7.

Simon Garrod and Martin J. Pickering. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8–11, 2004. doi: 10.1016/j.tics.2003.10.016. URL https://doi.org/10.1016/j.tics.2003.10.016.

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

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8048–8057. International Joint Conferences on Artificial Intelligence Organization, 2024. doi: 10.24963/ijcai.2024/890. URL https://doi.org/10.24963/ijcai.2024/890. Survey Track.

Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhuo Xu. LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578v3*, 2024a. doi: 10.48550/arXiv.2402.03578. URL https://arxiv.org/abs/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, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2146–2156. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/70222949cc0db89ab32c9969754d4758-Abstract.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (Natural) language. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hk8N3Sclg.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "Mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*, 1:9, 2024. doi: 10.1007/s44336-024-00009-2. URL https://link.springer.com/article/10.1007/s44336-024-00009-2.

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. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1495–1502, 2018. doi: 10.1609/aaai.v32i1.11492. URL https://ojs.aaai.org/index.php/AAAI/article/view/11492.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 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, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Zhengqi Pei, Qingming Huang, and Shuhui Wang. When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning. *arXiv preprint arXiv:2606.29354*, 2026. URL https://arxiv.org/abs/2606.29354.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249v11*, 2025. doi: 10.48550/arXiv.2501.14249. URL https://arxiv.org/abs/2501.14249v11.

Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004. doi: 10.1017/S0140525X04000056. URL https://doi.org/10.1017/S0140525X04000056.

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. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Linus Sander, Habtom Kahsay Gidey, Alexander Lenz, and Alois Knoll. A technical taxonomy of LLM agent communication protocols. *arXiv preprint arXiv:2606.19135*, 2026. URL https://arxiv.org/abs/2606.19135.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. Part I; Part II appears in the same volume, pp. 623–656, DOI 10.1002/j.1538-7305.1948.tb00917.x.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Reid G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, C-29(12):1104–1113, 1980. doi: 10.1109/TC.1980.1675516. URL https://doi.org/10.1109/TC.1980.1675516.

Elias Stengel-Eskin, Newton Sander, Carlos Bonetti, Sasha Boguraev, James Bowler, Hale Sirin, and Simon Kirby. GlossoGen: Emergent language in complex multi-agent LLM interactions. *arXiv preprint arXiv:2609.01491*, 2026. URL https://arxiv.org/abs/2609.01491.

Jianwen Sun, Yukang Feng, Yifan Chang, Chuanhao Li, Zizhen Li, Jiaxin Ai, Fanrui Zhang, Yu Dai, and Kaipeng Zhang. Closing the expression gap in LLM instructions via socratic questioning. *arXiv preprint arXiv:2510.27410*, 2025. URL https://arxiv.org/abs/2510.27410.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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

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

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. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. Beyond self-talk: A communication-centric survey of LLM-based multi-agent systems. *arXiv preprint arXiv:2502.14321*, 2025. doi: 10.48550/arXiv.2502.14321. URL https://arxiv.org/abs/2502.14321v2. arXiv:2502.14321v2.

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

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639v4*, 2025. doi: 10.48550/arXiv.2511.20639. URL https://arxiv.org/abs/2511.20639v4.

A Implementation Details

The main text treats CoGENT as a Self-coin–Share–Use communication layer around sequential debate. Table 8 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.

Table 8: Prompt components used by CoGENT and the vanilla baseline.

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 containing the opponent's answer and the detailed `KNOWN_TERMS` block. When the public board is nonempty, the prompt asks each debate response to explicitly reuse at least one visible term. Agent A sees the opponent's pre-debate answer, while Agent B sees Agent A's new answer from the same debate round.
P_{orig}	Vanilla debate prompt (Algorithm 2) containing the question, the agent's own previous answer, and the opponent's latest answer, without the `KNOWN_TERMS` block.

A.1 Verbatim Prompt Templates

Figure 7 lists the prompt templates used on FOLIO. Text in braces is filled in at run time.

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

Input: instance q, meta-rounds R
Output: final answer $\hat{a}$

initialize histories; $m_i^0 \leftarrow \text{LLMStep}(i, P_{\text{init}}(q))$ for $i \in \{A, B\}$
for $r = 1$ to R **do**
  $m_A^r \leftarrow \text{LLMStep}(A, P_{\text{orig}}(q, m_A^{r-1}, m_B^{r-1}))$
  $m_B^r \leftarrow \text{LLMStep}(B, P_{\text{orig}}(q, m_B^{r-1}, m_A^r))$
end for
$\hat{a} \leftarrow \text{Finalize}(m_A^R, m_B^R)$

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 Agent A's latest debate response and one prediction from Agent B'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

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

P_{init}: independent solving (Round 0)

```
Solve the logical reasoning problem from the premises and conclusion.
TRUE: the conclusion is forced. FALSE: its negation is forced. Otherwise: UNCERTAIN. Do not
    use outside knowledge or reverse one-way implications.
[PART 1] Current Turn's Answer
Explain briefly and check both whether the conclusion and its negation are forced.
[PART 3] Final Answer
End with exactly one of: #### TRUE  #### FALSE  #### UNCERTAIN
```

P_{judge} + P_{name}: private coining (Self-coin)

```
Your last response uses {word_count} words. Simplify it using new words and the KNOWN_TERMS
    currently visible to you. Write a NEW compression-only message. Do not solve the
    problem again. Do not output [PART 3], TRUE, FALSE, UNCERTAIN, or any #### line.
    Preserve proof status, polarity, quantifiers, negations, and one-way implication
    directions.
[PART 1] Compressed Reasoning / Term Usage
Use every term coined below.
[PART 2] New Words Coined in This Turn
new word = a brief definition
Leave blank if none.
```

P_{debate}: public debate (Use)

```
Here is another agent's solution:
{opponent_response}
Use it as additional information and provide updated reasoning and a final decision.
Problem:
{original_problem}
[PART 1] Current Turn's Answer
Give step-by-step logical analysis. If KNOWN_TERMS exist, reuse at least one. Check whether
    the conclusion and its negation are each forced.
[PART 3] Final Answer
End with exactly one of: #### TRUE  #### FALSE  #### UNCERTAIN
KNOWN_TERMS
{term = definition  invented_by: Agent X}
```

Figure 7: Prompt templates used on FOLIO.

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

$$B^{\text{van}}_n = \sum_{i\in\{A,B\}} \text{tok}(y^0_{i,n}),$$

$$F^{\text{van}}_n = \sum_{i\in\{A,B\}} \text{tok}(y^R_{i,n}).$$

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

$$B^{\text{cog}}_n = \sum_{i\in\{A,B\}} \text{tok}(y^0_{i,n}),$$

$$W^{\text{cog}}_n = \sum_{i\in\{A,B\}} \text{tok}(y^{\text{wu}}_{i,n}).$$

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

C Logged Case Traces and Author-Side Verification

This appendix collects the three logged cases discussed in the main text: FOLIO (Figure 1) and GPQA-Diamond and HLE (Figure 5). Figure 8 shows each case in full, with the problem statement, the independent VANILLA run, and the CoGENT run. Table 9 adds the independently run SINGLE baseline and the log keys. All controller outputs come from independent runs. Within each CoGENT column, entries follow the recorded turn order; they establish chronology and observed cross-speaker use, not the isolated causal effect of a term. The paragraphs after the table verify each gold answer independently of the logged outputs.

QUESTION	VANILLA DEBATE	COGENT
FOLIO		
Determine whether the conclusion logically follows from the premises. **Premises** 1. The Picuris Mountains are a mountain range in New Mexico or Texas. 2. Juan de Onate visited the Picuris Mountains. 3. The Harding Pegmatite Mine, located in the Picuris Mountains, was donated. 4. No mountain range in Texas has a donated mine. **Conclusion** The Harding Pegmatite Mine is not located in a mountain range in New Mexico.	**Selected answer: TRUE (entailed) ×** **Agent A:** If the conclusion is false, the Harding Pegmatite Mine is located in a mountain range in New Mexico. This is mistakenly treated as contradicting Premise 3. **Agent B:** Since the donated mine is in the Picuris Mountains, which are in New Mexico or Texas, the debate concludes that the statement is forced by the premises. Both agents preserve the same logical reversal: "The conclusion remains entailed by the premises." **Final consensus:** **TRUE (entailed) ×**	**Selected answer: FALSE (contradicted) ✓** **Round 0 - Initial answers** **Agent A:** "The conclusion is TRUE." **Agent B:** "The conclusion is FALSE given the premises." **Agent B coins novsure** = a situation where the conclusion contradicts the existing premises. **Share: novsure** is published to the Shared Term Board and becomes visible to Agent A. **Agent B:** Premise 3 places the donated mine in the Picuris Mountains. Premise 1 places that range in New Mexico or Texas, while Premise 4 rules out Texas. **Agent A reuses novsure:** The mine must therefore be in New Mexico. Since the conclusion states that it is *not* in New Mexico, the conclusion is a **novsure**. **Agent A checks:** The shared definition agrees with Premises 1, 3, and 4, so Agent A revises its initial TRUE answer. **Agent B confirms:** The conclusion remains contradicted. **Final consensus:** **FALSE (contradicted) ✓**
GPQA-DIAMOND		
Based on the following matrices, choose the correct statement. $W = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ $X = \begin{bmatrix} i & -1 & 2i \\ 1 & 0 & 1 \\ 2i & -1 & -i \end{bmatrix}$ $Y = \begin{bmatrix} 0.5 & 0.1 & 0.2 \\ 0.1 & 0.25 & 0.1 \\ 0.2 & 0.1 & 0.25 \end{bmatrix}$	**Selected answer: Option C ×** **Option C:** e^X changes the norm of at least one vector. **Agent A:** $(X^\dagger)_{3,2} = -1$, but $(-X)_{3,2} = +1$. Thus X is treated as non-anti-Hermitian, so e^X is judged non-unitary. **Agent B:** Since e^X is not unitary, the similarity transform is said not to preserve Hermiticity, and at least one vector is assumed to change norm. **Vanilla Debate:** Both agents independently repeat the same conclusion. The key chain is stated as: e^X preserves all norms $\Leftrightarrow$ e^X is unitary $\Leftrightarrow$ X is anti-Hermitian. Because the entry at (3, 2) is misread, the chain is incorrectly broken and Option C is retained. **Final consensus:** **Option C ×**	**Selected answer: Option B ✓** **Option B:** $e^X Y e^{-X}$ represents a quantum state. **Round 0 disagreement:** Agent A selects B; Agent B selects C. **Agent A coins zelvorn** = the anti-Hermitian property $M^\dagger = -M$. **Share: zelvorn** is published to the Shared Term Board before Agent B audits the disputed entries. **Agent B uses the shared term to check:** $(X^\dagger)_{3,2} = \mathrm{conj}(X_{2,3}) = \mathrm{conj}(1) = 1 = (-X)_{3,2}$ ✓ $(X^\dagger)_{1,2} = \mathrm{conj}(X_{2,1}) = \mathrm{conj}(1) = 1 = (-X)_{1,2}$ ✓ $(X^\dagger)_{1,3} = \mathrm{conj}(X_{3,1}) = \mathrm{conj}(2i) = -2i = (-X)_{1,3}$ ✓ The diagonal entries i, 0, $-i$ are purely imaginary or zero and also satisfy $X^\dagger = -X$. Therefore, X is **zelvorn**. Hence, e^X is unitary and $(e^X)^\dagger = \exp(X^\dagger) = e^{-X}$. **Agent B corrects:** Because the shared audit proves unitarity, e^X preserves the norm of every vector, so Option C is false. **Shared conclusion:** Y is Hermitian, positive semidefinite, and trace one. Unitary conjugation preserves these properties, so $e^X Y e^{-X}$ remains a valid density matrix, explicitly validating Option B. **Final consensus:** **Option B ✓**
HLE		
Bootstrap Percolation A disease spreads across an $n \times n$ grid. Once a site is infected, it remains infected. At each step, a healthy site becomes infected if at least three of its neighbors are infected. For $n = 14$, what is the minimum number of initially infected sites needed to eventually infect the entire grid? **Required answer: an exact integer.**	**Selected answer: 75 (candidate) ×** Since every corner has degree 2, all four corners must initially be infected. A boundary-counting argument gives the repeated candidate $\lceil (n^2 + 2n) / 3 \rceil = \lceil 224 / 3 \rceil = 75$. Intermediate turns repeatedly produce Answer 75, but the last response ends in the middle of a formula. Vanilla Debate treats a lower bound as the exact minimum and does not complete a valid answer-bearing consensus. **Final consensus:** **No valid consensus ×**	**Selected answer: 76 ✓** **Round 0:** Agent A proposes 75; Agent B's response is truncated before a valid Answer line. **Agent B coins evengridform for the exact even-grid case:** $\mathbf{evengridform}(n) = \lceil (n^2 + 2n + 4) / 3 \rceil$, for even n. **Share: evengridform** enters the Shared Term Board for Agent A. **Agent A reuses evengridform:** $n = 14$ is even (and $14 \equiv 2 \pmod 6$), so $\lceil (196 + 28 + 4) / 3 \rceil = \lceil 228 / 3 \rceil = 76$. The **+4** correction raises the lower bound 75 to the even-grid value 76. Thus, the minimum number of initially infected sites required to infect the entire grid is 76. Both agents preserve and verify the corrected formula. An explicit Answer line repairs the extraction failure. **Unanimous answer-bearing consensus:** Agent A: 76; Agent B: 76. **Final consensus:** **76 ✓**

Figure 8: Full panels for the three logged cases. Each row gives the problem statement, the independent VANILLA run, and the COGENT run; coined terms are shown in green. Panels are condensed from the logs, and the COGENT column follows the recorded turn order. Log keys and the SINGLE baseline are in Table 9.

Table 9: Audit summary for the three cases in Figure 8. Baseline and CoGENT outputs come from independent runs. Internal indices and log keys refer to the supplied trace files; they are not benchmark-defined item IDs. The CoGENT chronology records turn order and observed cross-speaker use without claiming a term-level causal effect.

Case and log key	Independent baselines	Logged CoGENT chronology
FOLIO, validation index 114 (zero-based)	SINGLE: TRUE. VANILLA: both agents remain at TRUE.	Round 0 splits: Agent A gives TRUE, while Agent B gives FALSE. During *Self-coin*, Agent B creates `novsure`. *Share* places the accepted term on the public board. During *Use*, Agent A applies the term and changes its answer to FALSE. Debate 1 reaches FALSE/FALSE, and later rounds maintain that answer. `novcontra` appears only after this consensus.
GPQA-Diamond, log 178	SINGLE: C. VANILLA: both agents remain at C.	Round 0 splits: Agent A gives B, while Agent B gives C. During *Self-coin*, Agent A creates `zelvorn` for anti-Hermiticity. *Share* makes the definition visible to Agent B. During *Use*, Agent B rechecks $(X^\dagger)_{32} = \overline{X_{23}} = 1 = (-X)_{32}$ and changes its answer from C to B. The final consensus is B/B.
HLE, log 02	SINGLE: the output is truncated before a valid answer line. VANILLA: intermediate turns repeatedly give 75, but the run ends without a valid unanimous answer.	In Round 0, Agent A proposes 75, while Agent B's output ends before a valid answer line. During *Self-coin*, Agent B creates `evengridform` for the exact even-grid formula. *Share* adds the term to the public board. During *Use*, Agent A applies it at $n = 14$ and obtains 76. Both agents keep the corrected formula and finish with a 76/76 consensus.

FOLIO proof kernel. The Harding Pegmatite Mine is a donated mine located in the Picuris Mountains, and the Picuris Mountains are a mountain range in New Mexico or Texas. If the Picuris Mountains were in Texas, a Texas mountain range would contain a donated mine, contradicting the premises. They must therefore be in New Mexico. The Harding Pegmatite Mine is thus located in a mountain range in New Mexico, so the negative conclusion is FALSE. The Juan de Oñate premise is not needed for this label.

GPQA-Diamond predicate check. Direct conjugate-transposition of the matrix X in Figure 8 gives $X^\dagger = -X$. In particular, the disputed entry satisfies $(X^\dagger)_{32} = \overline{X_{23}} = 1 = (-X)_{32}$; the contrary value in the initial C response is an indexing error. Anti-Hermiticity implies that $U = e^X$ is unitary and $e^{-X} = U^\dagger$; consequently $e^X Y e^{-X} = UYU^\dagger$ preserves a valid density matrix. The logged supporting checks are $Y = Y^\dagger$, $\mathrm{Tr}(Y) = 1$, and positive leading minors 0.5, 0.115, and 0.01775.

HLE exact-value check. Let $m(G, 3)$ denote the minimum size of an initial set that infects all vertices of G under the three-neighbor rule. The published result gives the following exact value for every even n (Benevides et al., 2024):

$$m(P_n \square P_n, 3) = \left\lceil \frac{n^2 + 2n + 4}{3} \right\rceil .$$

Since 14 is even,

$$m(P_{14} \square P_{14}, 3) = \left\lceil \frac{14^2 + 2 \cdot 14 + 4}{3} \right\rceil = 76.$$

By contrast,

$$\left\lceil \frac{14^2 + 2 \cdot 14}{3} \right\rceil = 75$$

omits the $+4$ correction and is one below the exact minimum. It does not establish the exact value. This calculation is an author-side verification of the logged answer and is not attributed to `evengridform` itself.

Table 10: Design of the component-focused MMLU suite. Vanilla Debate and Full CoGENT are anchors; the six intermediate conditions test call-budget alternatives or remove one proposed lexical mechanism.

Condition	Controlled intervention
Vanilla Debate (anchor)	Uses two initial calls and three ordinary debate/revision pairs (eight calls total). It has no Self-coin, Share, or term-guided Use stage.
Call-Matched Debate	Adds ordinary revision calls to match Full CoGENT's 16-call budget. It has no task-local terms, definitions, term board, or term-reuse rule, and tests whether additional communication alone is sufficient.
Summary-Only	Replaces Self-coin with private natural-language summaries. It creates no named terms and has no Share stage or term-guided Use, isolating ordinary private-state compression.
Injected Lexicon	Removes Self-coin and instead gives both agents the same fixed lexicon with definitions. This tests whether a pre-existing vocabulary can replace terms created for the current problem.
CoGENT-NoDef	Keeps dynamic term creation, Share, and the later-reuse rule, but removes the definition component of Self-coin and publishes only opaque handles. When the public board is nonempty, each debate prompt asks for at least one visible handle to be reused.
CoGENT-Private	Keeps dynamic term creation, same-response use, and the later-reuse rule, but removes Share: each agent sees only its own term board. When that private board is nonempty, the debate prompt asks the agent to reuse at least one of its own terms. This tests private shorthand without cross-agent visibility.
CoGENT-Shared-Optional	Keeps Self-coin and Share, but makes the term-reuse part of Use optional: a debate response may be completed without reusing any visible term.
Full CoGENT (anchor)	Uses the complete 16-call Self-coin–Share–Use protocol. Self-coin creates definition-bound terms and uses them in the same response; Share places accepted terms on the public board; when that board is nonempty, each Use prompt asks the debate response to reuse at least one visible term.

Table 11: Reported GPT-3.5-Turbo MMLU component-suite results. Accuracy uses the denominator in "Correct / eval."; CR is the per-example compression rate, where lower is stronger retained-state compression, and final tokens measure post-warm-up debate output rather than complete API usage. Full CoGENT has the highest reported accuracy and strongest final-stage compression, while its total-token cost is the largest.

Condition	Correct / eval.	Acc. (%)	CR (%, mean ± SD)	Final tokens (mean ± SD)	Total tokens
Vanilla Debate	342 / 500	68.40	95.15 ± 24.95	361.838 ± 93.953	6,571,922
Call-Matched Debate	364 / 500	72.80	90.30 ± 24.86	296.840 ± 87.619	7,167,772
Summary-Only	362 / 500	72.40	60.79 ± 22.00	193.314 ± 48.734	4,787,217
Injected Lexicon	353 / 500	70.60	80.00 ± 29.06	259.222 ± 77.373	7,014,334
CoGENT-NoDef	323 / 458	70.52	71.68 ± 24.57	230.362 ± 69.770	6,376,632
CoGENT-Private	335 / 482	69.50	69.09 ± 23.65	223.706 ± 60.168	6,236,944
CoGENT-Shared-Optional	331 / 480	68.96	73.46 ± 25.61	235.816 ± 62.239	6,186,622
Full CoGENT	376 / 500	**75.20**	$\mathbf{37.27 \pm 16.83}$	$\mathbf{107.292 \pm 23.951}$	16,707,012

D MMLU Ablation Details

D.1 Overall Results of Ablation Study

Tables 10 and 11 in Appendix D report six controls on the same 500-item GPT-3.5-Turbo MMLU sample. Full CoGENT has the highest accuracy at 75.20%, 6.80 points above the Vanilla anchor (68.40%); Call-Matched Debate and Summary-Only reach 72.80% and 72.40%, so Full stays above both additional communication and ordinary self-compression. Injecting a frozen shared lexicon reaches 70.60%, and the component removals give NoDef 70.52%, Private 69.50%, and Shared-Optional 68.96%, an ordering consistent with explicit definitions, cross-agent visibility, and enforced later uptake acting jointly rather than as interchangeable add-ons. Full also gives the strongest final-stage compression—a 37.27% compression rate and 107.29 final-stage tokens against the anchor's 95.15% and 361.84, a 70.35% reduction—while its total token cost stays higher (16.71M versus 4.79M–7.17M). The controls and anchors differ in prompt and decoding configuration, and three conditions use analysis-eligible subsets, so the suite isolates components rather than matching them exactly. Within those constraints the ordering is strictly monotone across all eight conditions.

Tables 10 and 11 report the complete component-focused GPT-3.5-Turbo MMLU suite summarized in the main text. Every condition uses the same 500 sampled item identifiers (sample seed 42). Vanilla Debate and Full CoGENT serve as anchors. The six intermediate conditions distinguish call-budget-matched ordinary communication and generic compression from lexical mechanisms, then remove or replace one proposed part of the full protocol at a time.

Evaluation and completeness notes. The six intermediate controls were executed with the deterministic `causal10_v1` configuration (temperature 0.0), whereas the Vanilla and Full anchors are existing `frne` benchmark artifacts (temperature 0.7). The 68.40% Vanilla anchor is therefore the later `frne` suite run (also reported in Table 1), not an earlier 64.80% Vanilla run; Full is the consolidated 500-item artifact used for the reported 75.20% result. The rows share sampled items but are not a strictly call-, prompt-, and decoding-matched causal comparison. NoDef, Private, and Shared-Optional contain 42, 18, and 20 analysis-ineligible incomplete examples, respectively; their reported accuracies use 458, 482, and 480 evaluated items. Summary-Only combines its interrupted first shard with the 285-item resume by example identifier after deduplication. These differences motivate the descriptive language in the main text and the complete rerun requirement in the Limitations section.

Within those constraints, the component ordering is informative. Full CoGENT is 2.40 points above Call-Matched Debate, 2.80 above Summary-Only, and 4.60 above Injected Lexicon. The reported gaps associated with NoDef, Private, and Shared-Optional are 4.68, 5.70, and 6.24 points, respectively, corresponding to the definition component of Self-coin, cross-agent Share, and prompted later reuse in Use. Full also reduces final-stage token count by 70.35% relative to the suite's Vanilla anchor. However, its 16.71M total tokens exceed the 4.79M–7.17M range of the intermediate controls, so this efficiency result concerns the final communication stage rather than total API use.

E Lexical-Function Annotation Protocol

For the soft lexical-function analysis summarized in Figure 4, we use the same six-category codebook across old and new datasets: discriminative cue, domain concept, quantity/unit variable, operation checkpoint, proof-status logic, and gap/uncertainty marker. The original MMLU–FOLIO–GSM8K block was annotated through human annotation followed by LLM-assisted verification and manual adjudication. For the added MedQA–GPQA-Diamond–HLE block, we reuse the six categories but extend the adjudication guide for domain-specific ambiguity: for example, a medical *condition* is a domain concept unless it functions as a logical constraint; an *energy gap* or *anion gap* is a quantity or domain variable rather than an evidence gap; and a *threshold signature* is a cryptographic domain concept rather than a numeric threshold. A verifier is given the term, definition, task family, and candidate category and asked to flag unsupported or ambiguous labels. Flagged cases are manually inspected. Percentages are normalized over problem-term counts within each task and model.

F Additional Lexical Ecology Notes

Tables 4 and 5 in the main text report the split lexical-ecology analysis. Table 4 separates accepted entries, within-task unique surface terms, and shared/public availability, while Table 5 preserves representative logged term–definition entries and their criterion-level interpretation. Appendix E gives the lexical-function annotation protocol, and Appendix G reports the full Werewolf lexical-grounding diagnostics.

G Additional Werewolf Lexical-Grounding Metrics

Table 6 in the main text reports the compact original five-game Werewolf signals, while Table 7 reports the 50-seed visibility-ablation mechanism check. Tables 12, 13, and 14 retain the auxiliary function, comparison, footprint, productivity, and persistence diagnostics for the original stress test. Table 15 gives the fuller GPT-4o-mini ablation diagnostics, including the outcome row and the exact public-definition leakage check reported for the ablation condition.

G.1 Visibility Ablation and Grounded Atomic Uptake Density

A 50-seed GPT-4o-mini visibility ablation makes the mechanism explicit (Table 7). The faction-private condition accepts nearly as many terms as full CoGENT (109 vs. 114), yet observed reuse falls from 776 to 337, lexical uptake from 6.81 to 3.09, public reuse share from 60.57% to 27.89%,

Table 12: Functional ecology of accepted Werewolf terms.

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 13: Werewolf comparison with ORIGIN/no-COGENT.

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.

and the aggregate public-uptake advantage from $1.71\times$ to $0.89\times$. The contrast is therefore not a difference in lexical volume.

To compare uptake under unequal trajectory lengths, we model each accepted definition-bound term as a candidate handle for a task-local reasoning atom and define grounded atomic uptake density as $\mathrm{GAUD}(B) = \sum_g R_g / \sum_g B_g$, where R_g is protocol-observed exact-term reuse in game g and B_g is a communication budget. Normalizing by complete-trajectory tokens and by model calls both preserve the advantage ($2.00\times$ and $2.07\times$; paired-seed bootstrap 95% intervals $[1.04, 3.64]$ and $[1.07, 3.83]$), so the Werewolf result is better explained by public and shared uptake than by explicit coinage, private lexical storage, longer trajectories, or additional calls. GAUD measures observable handle deployment, not semantic understanding, and is deliberately an operational lower bound: one definition may encode several semantic relations, while a term is counted at most once per response. Appendix G reports the full diagnostics and the villager-win outcomes, which favor the ablation and are not part of this claim.

Table 14: Full exploratory Werewolf lexical-grounding metrics.

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×	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.

H GENERATIVE AI STATEMENT

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

Table 15: Full Werewolf visibility-ablation diagnostics on 50 GPT-4o-mini games, seeds 42–91. The table compares the full CoGENT Werewolf run against the faction-private ablation. Outcome is included for transparency, but the mechanism claim in the main text is based on lexical-grounding diagnostics rather than villager-win rate.

Metric	Full CoGENT	Faction-private ablation	Interpretation
Villager-win outcome	26 / 50	37 / 50	Outcome is reported separately from lexical grounding; NORMAL reaches 34 / 50 in the same 50-seed batch.
Accepted Werewolf terms	114	109	Term volume is comparable, so the main contrast is not lexical quantity.
Observed term reuses	776	337	Ablation substantially reduces recorded uptake.
Lexical uptake rate	6.81	3.09	Reuse per accepted term drops under faction-private visibility.
Final lexicon snapshot tokens	2028	1836	Stored lexicon size remains compact and close to full CoGENT.
Lexicon footprint ratio	0.112	0.110	Footprint is nearly unchanged; productivity, not storage size, changes.
Reuse per 100 lexicon tokens	38.26	18.36	Ablation is less productive per stored lexicon token.
Public terms / reuses	54 / 470	33 / 94	Public-pool reuse falls sharply.
Wolf-private terms / reuses	60 / 306	76 / 243	Wolf-private reuse becomes the dominant ablation channel.
Public reuse share	60.57%	27.89%	Public terms no longer dominate reuse under ablation.
Aggregate public uptake advantage	1.71×	0.89×	Public terms are not more productive per entry than wolf-private terms in the ablation aggregate.
Terms used in public speech	13 / 114	13 / 109	The same absolute count surfaces in public speech, but with much lower public reuse.
Terms used by ≥ 2 public speakers	2 / 114	3 / 109	Cross-speaker public uptake remains sparse in both GPT-4o-mini 50-seed batches.
Public checkpoint persistence	0.081	0.076	Persistence is slightly lower under ablation.
Games with any accepted terms	34 / 50	32 / 50	Term formation is similarly frequent across games.
Games with any observed reuse	32 / 50	30 / 50	Reuse appears in slightly fewer ablation games.
Public definition exact leakage to wolf logs	Not reported	0	Exact-string scan detects no public-definition leakage into wolf personal logs in the ablation condition.