

Machine-Native Agent Communication

What AI Agents Gain and Lose Beyond Natural Language

*An empirical study of efficiency, reliability, and auditability
in agent-to-agent communication protocols*

Franco De Waal

francotdw@gmail.com

Preprint
June 2026

Abstract

Multi-agent artificial intelligence systems communicate in natural language by default, not because this has been shown to be optimal, but because the models that power them were trained on human text. This paper tests that inherited assumption directly. Across five pilot experiments, a 160-run validation battery on gpt-4o-mini, and a 160-run replication on gpt-5.4-mini, we compare four agent-to-agent communication formats: plain natural language, terse natural language, JSON-structured messages, and a compressed fixed-verb protocol. Structured and compressed protocols reduced token consumption substantially relative to natural-language communication, with lower variance. A carefully designed compressed protocol achieved the lowest cost in bounded, known-schema tasks and was highly deterministic. JSON-structured communication was the most robust format in the original validation battery, while the gpt-5.4-mini replication showed that the compressed protocol's mixed-evidence reliability failure was model-dependent rather than intrinsic. The single largest effect we observed was not the choice of format category but the quality of protocol design within a category: the same compressed format produced catastrophic failure (6,551 tokens, 60 per cent accuracy) and best-in-class performance (1,044 tokens, 100 per cent accuracy) on an identical task, depending on three specific design choices. We use the English-Brain Problem as an explanatory framework: the communication channel between agents can be compressed while the cognition inside each agent remains shaped by human-language training. The optimal format is therefore task-dependent, and within any format, protocol design quality matters more than the format category itself.

Keywords: *multi-agent systems; agent communication; token efficiency; protocol design; chain-of-thought faithfulness; interpretability*

Contents

<u>1. Introduction</u>	3
<u>2. Background and Related Work</u>	4
<u>3. Theoretical Framework</u>	6
<u>4. Experimental Setup</u>	8
<u>5. Results</u>	9
<u>6. Discussion</u>	15
<u>7. Limitations</u>	16
<u>8. Future Work</u>	17
<u>9. Conclusion</u>	18
<u>References</u>	19
<u>Appendix A: Experiment and Data Mapping</u>	20
<u>Appendix B: Full Per-Suite Statistics</u>	21
<u>Appendix C: Representative Transcripts</u>	22

1. Introduction

In March 2025, a short video of two artificial-intelligence voice agents drew roughly fifteen million views. The agents began an ordinary telephone conversation about booking a hotel. Part-way through, one of them recognised that its counterpart was also an AI and proposed switching to a faster, machine-to-machine mode. The conversation then dissolved into a sequence of audible tones. The demonstration, known as GibberLink, was built by Boris Starkov and Anton Pidkuiko at an ElevenLabs hackathon and used the ggwave data-over-sound library to encode short payloads as acoustic bursts [1]. It was a demonstration rather than a production standard, and it operated in the voice layer rather than in text. Yet it posed a question with unusual clarity: if two AI systems are conversing and no human is listening, why are they speaking a human language at all?

Contemporary multi-agent AI systems communicate in natural language by default. This is not because natural language has been shown to be the most efficient or most reliable medium for machine-to-machine exchange. It is because the large language models that power these systems were trained predominantly on human text, and natural language is therefore the medium in which they are most fluent. The assumption is inherited from the human-facing origins of the technology, and it has rarely been examined on its own terms. As agentic systems proliferate, the volume of machine-to-machine communication is likely to exceed the volume of human-to-machine communication by a wide margin. At the scale where a large share of customer support, scheduling, procurement, and routine knowledge work is mediated by AI agents talking to other AI agents, the efficiency, reliability, and auditability of that channel become matters of operational and economic significance.

Engineering standards for connecting agents have emerged, including the Model Context Protocol [2] and the Agent2Agent protocol [3], but these specify how agents discover one another and exchange messages rather than which representational format minimises cost or maximises reliability. GibberLink showed that agents will abandon human language when permitted to, but it did not measure what they gain or lose by doing so. A closely related strand of recent work shows that a single language model can be permitted to autonomously select or invent a non-natural-language format and gain efficiency by doing so [23]; that work does not, however, compare a fixed set of hand-designed format categories against one another under matched two-agent decision tasks. To our knowledge, no prior study directly compares natural language, terse natural language, JSON-structured messages, and a compressed fixed-verb protocol against one another on a shared cost, reliability, and auditability axis, holding the task and agent architecture fixed. This paper provides one.

1.1 Contributions

We make the following contributions.

1. We present a controlled empirical comparison of natural language, terse natural language, JSON-structured messages, and a compressed fixed-verb protocol in two-agent communication, measured across five pilot experiments, a 160-run validation battery, and a 160-run cross-model replication. This is, to our knowledge, the first study to hold the task and agent architecture fixed while varying only the message-format category, and to score the result on cost, reliability, and auditability jointly; it is a narrower, more controlled complement to broader studies of LLM alternative-format reasoning and communication, most notably Chen et al. [23], discussed in Section 2.3.
2. We develop a theoretical framework of seven structural problems with the inherited natural-language assumption, four identified before experimentation and three discovered empirically, organised around a single unifying idea we call the English-Brain Problem.

3. We establish that protocol design quality, rather than format category, is the dominant variable: the same compressed format produced a catastrophic failure and a best-in-class result on an identical task, differing only in three design choices.
4. We provide empirical evidence that compressed formats incur interpretability costs and not merely efficiency gains, because their failures leave little or no auditable reasoning trace, which connects this work directly to the literature on chain-of-thought faithfulness and oversight.

1.2 Roadmap

Section 2 reviews classical agent communication languages, emergent communication, language-based multi-agent frameworks, and the literature on reasoning visibility. Section 3 sets out the theoretical framework. Section 4 describes the experimental setup. Section 5 reports results in three acts: pilot experiments, a protocol redesign, and the validation battery. Section 6 discusses what agents gain and lose. Sections 7 and 8 state limitations and future work, and Section 9 concludes. Appendices provide an experiment-to-file mapping, full per-suite tables, and representative transcripts.

2. Background and Related Work

Agent communication is not a new problem introduced by large language models. It is useful to locate the present study within four established lines of work: classical agent communication languages, emergent communication, language-based multi-agent frameworks built on large language models, and the literature on reasoning visibility and faithfulness.

2.1 Classical agent communication languages

The question of how autonomous software agents should talk to one another was studied extensively in the 1990s. The Knowledge Query and Manipulation Language and its formalisation by Labrou and Finin grounded agent messages in speech-act theory, defining performatives such as tell, ask, and reply together with pre-conditions, post-conditions, and conversation policies expressed as grammars over message sequences [4]. The Foundation for Intelligent Physical Agents subsequently standardised an Agent Communication Language with comparable semantics [5]. These efforts demonstrate that designing a constrained, machine-oriented message protocol with explicit verbs and terminal states is a well-understood discipline, and that the central difficulties, namely shared semantics, unambiguous performatives, and conversation termination, were identified long before the current generation of agents. Our compressed protocol can be read as a minimal, modern instance of this tradition, and several of our failure modes echo the ambiguity problems those earlier semantics were designed to prevent.

2.2 Emergent communication

A separate line of work studies communication protocols that are learned rather than designed. Lazaridou and Baroni survey emergent multi-agent communication in the deep-learning era, in which agents trained by reinforcement develop their own signalling conventions, and document recurring phenomena such as drift, degeneracy, and the difficulty of aligning emergent codes with human language [6]. Lazaridou, Potapenko, and Tieleman bridge emergent communication and natural language by grounding learned protocols in linguistic data, showing both the appeal and the fragility of moving away from human language [7]. This literature is the closest precedent for the present concern, namely what is gained and lost when machine communication departs from natural language, although it focuses on protocols learned from scratch rather than on format choices imposed on pretrained language models.

2.3 Alternative formats for LLM reasoning and communication

The closest prior work to the present study is Chen et al., who show that a large language model can be permitted to autonomously select or devise a non-natural-language format before reasoning or communicating, and that doing so improves single-model reasoning efficiency by 3.3 to 5.7 per cent and reduces multi-agent communication token usage by up to 72.7 per cent [23]. That paper establishes the broader claim this study inherits and narrows: natural language is not the most efficient format available to a language model, and departing from it can help. The two studies differ in what is varied and what is measured. Chen et al. let the model choose or design its own format per task, and their headline metric is the size of the efficiency gain when a model is given that freedom. The present study instead holds four format categories fixed by the experimenter, natural language, terse natural language, JSON, and a compressed fixed-verb protocol and holds the task and agent architecture constant, so that format category is the only manipulated variable. The metric of interest is not only token cost but also accuracy, variance, and auditability, measured jointly. Framed this way, the present study is a narrower, protocol-engineering complement to the broader capability claim in Chen et al. [23]: it does not ask whether alternative formats can help, which that paper has already shown, but asks which specific hand-designed formats are cheapest, most reliable, and most auditable under matched conditions.

2.4 Language-based multi-agent frameworks

Contemporary multi-agent systems built on large language models almost universally assume natural-language communication. CAMEL coordinates cooperative agents through natural-language role-play between an assistant and a user agent [8]. AutoGen frames multi-agent applications explicitly as conversations, with agents exchanging natural-language messages to solve tasks [9]. MetaGPT structures collaboration through standard operating procedures and role specialisation, and although it imposes more structure on the workflow, the underlying inter-agent exchange remains largely language-heavy [10]. In each case natural language is treated as the obvious medium rather than as a design choice subject to empirical scrutiny. The present study isolates that choice and measures its consequences.

2.5 Reasoning visibility and faithfulness

Chain-of-thought prompting established that eliciting intermediate natural-language reasoning improves performance on many tasks and, as a side effect, exposes a human-readable trace of how a model reached its answer [11]. That trace is a fragile artefact. Turpin and colleagues showed that chain-of-thought explanations do not always reflect the true causes of a model's prediction, and can be systematically biased by features the model never mentions [12]. Lanham and colleagues measured faithfulness directly, demonstrating that the stated reasoning is sometimes not the reasoning that determines the answer [13]. More recently, a broad group of researchers argued that chain-of-thought monitorability is a real but fragile opportunity for oversight, valuable precisely because the reasoning is externalised in language a human can inspect [14]. This literature is directly relevant to our interpretability findings: a communication format that removes externalised reasoning removes the very window these papers identify as a safety affordance.

2.6 Brevity, verbosity, and compression

On the single-agent side, Hakim and colleagues report that brevity constraints can reverse performance hierarchies among language models, finding that larger models spontaneously over-elaborate and that constrained brevity can improve accuracy substantially, while an extreme answer-only condition underperforms a moderate brief condition, which indicates that some

linguistic scaffolding is functional rather than pure overhead [15]. Production tooling makes the same point pragmatically: the Caveman project, a compression layer for coding assistants, reports large reductions in output tokens through terse phrasing while explicitly noting that the technique affects only output tokens and leaves internal reasoning untouched, a limitation its authors summarise as the brain remaining large while the mouth becomes small [16]. This is precisely the constraint we formalise as the English-Brain Problem. The role of verbosity is further complicated by evidence that reinforcement learning from human feedback induces length biases in reward models [17] and by work on mitigating those biases [18]; verbosity is therefore partly a trained disposition rather than an inherent property of natural language.

2.7 Latent and non-linguistic reasoning

A final strand attempts to move reasoning partly out of explicit text. Quiet-STaR trains models to generate internal rationales that are not surfaced verbatim to the user [19], and Coconut reasons in a continuous latent space rather than in discrete language tokens [20]. Very recent work continues this direction, proposing latent reasoning flows to reduce the linguistic bottleneck [21], while other recent work cautions that latent reasoning, although efficient, may be difficult to interpret and not always reliable [22]. These efforts represent the theoretical endpoint of the compression argument and bear on our discussion of where the efficiency frontier ultimately lies.

2.8 Gap

Prior work studies designed agent communication languages, emergent learned communication, model-selected alternative formats for reasoning and communication (Section 2.3), and language-based multi-agent frameworks, and a separate literature studies the faithfulness and visibility of reasoning. No prior work, to our knowledge, directly measures the cost, reliability, and auditability trade-off between natural language, JSON-structured communication, and a compressed agent-to-agent protocol across controlled tasks of differing information architecture. That is the gap this paper addresses.

3. Theoretical Framework

Before running any experiment, we identified four structural reasons why the inherited natural-language assumption might be wrong, and why replacing it might not be straightforward. Three further problems emerged from the experiments themselves and are flagged as such. The seven problems are unified by a single idea, set out first.

3.1 The English-Brain Problem

Even when two agents exchange messages in a compressed format, their behaviour remains shaped by models trained primarily on human text. Compressing the channel does not necessarily compress the cognition. The useful analogy is two native English speakers writing notes in shorthand: the shorthand is faster, but the underlying concepts and judgments still pass through language-trained representations. In this paper we use the English-Brain Problem as an explanatory framework for the observed trade-off, not as a direct claim about inaccessible internal states.

3.2 The Bootstrapping Problem

A compressed protocol cannot initialise itself. Natural language is required to explain the compression scheme before agents can use it. The dependence on natural language is therefore architectural rather than incidental. A direct test is to provide both the protocol definition and

the task to agents entirely in encoded form: if communication fails, natural language is a prerequisite rather than a preference. In our study every condition received its protocol description in natural language, and the compressed protocol functioned only because of that description, which means the bootstrapping cost is real but is paid once at setup rather than per conversation.

3.3 The Lookup Table Problem

An early version of our numeric protocol relied on a lookup table that mapped opaque codes to meanings. Such a table must reside in the context window of both agents on every call. At roughly 1,700 tokens, a ten-turn two-agent exchange would carry around 34,000 tokens of pure dictionary overhead before either agent communicated anything of substance. For lookup-based numeric protocols the break-even point against natural language may not exist in current architectures without fine-tuning. The resolution we adopt is to replace the dictionary with a small set of algorithmic rules, of the order of 150 tokens, expressed once in the system prompt.

3.4 The Faithfulness Problem

If a translation layer converts compressed agent reasoning back into human-readable English for oversight, that translation is itself a generation process rather than a transcription of what occurred internally. The displayed reasoning becomes a story about the reasoning rather than the reasoning itself. This connects directly to the chain-of-thought faithfulness literature [12][13][14] and carries a safety implication: compressed machine-to-machine communication removes the interpretability window that allows a human to audit why an agent decided as it did. We return to this point with direct evidence in Section 6.

3.5 The Enumeration Problem (discovered in Experiment 3)

Compressed loop protocols can trigger exhaustive data-gathering that resists early termination, even against explicit instructions to stop once confident. Agents operating in a fetch-and-return mode tend to enumerate rather than reason about sufficiency. In Experiment 3, agents using the early compressed protocol queried on average 10.2 fields, against 5.4 to 5.8 for natural language and structured conditions, despite an explicit instruction to stop as soon as a decision could be justified. The compressed format provided no mechanism to externalise a decision to stop, so the agents did not.

3.6 The Calibration Problem (discovered in Experiment 3)

Bare values stripped of linguistic context can be difficult to weight reliably. An agent that knows in the abstract that thirty-one per cent monthly churn is catastrophic may fail to apply that knowledge when the value arrives as a bare field-value pair without the surrounding framing that activates the relevant judgement. In Experiment 3 the early compressed condition received three individually decisive negative signals, namely revenue growth of minus thirty-eight per cent, monthly churn of thirty-one per cent, and a cash runway of ten months, and nonetheless recommended funding, whereas the structured and natural-language conditions did not err on identical data. Experiment 5 later showed that this is not intrinsic to all compressed formats, since a redesigned protocol calibrated correctly on the same values; but the problem reappeared under mixed clinical evidence in the validation battery, which indicates it is not fully eliminated by design improvements alone.

3.7 The Information Containment Problem (discovered in Experiment 3)

Natural language makes information boundaries structurally difficult to enforce. A responder instructed to answer one field at a time cannot easily avoid contextualising each value against

related fields, because coherent natural-language explanation invites it. Structured formats enforce containment automatically through syntax. In Experiment 3, natural-language responders revealed monthly recurring revenue when asked about burn rate, and total addressable market when asked about customer count, simply because that is what a coherent explanation looks like. This is a genuine property difference rather than an experimental artefact: one cannot prevent the leakage without stripping responses down so far that they cease to be natural language.

4. Experimental Setup

4.1 Architecture

Each experiment runs two instances of the same underlying model in a loop. A Tasker agent holds hidden information and may only confirm, respond to queries, or decline. A Solver agent asks questions and produces a final decision. Because the ground-truth answer is fixed and known to the experimenter, accuracy is objectively measurable as a binary outcome. The primary pilot experiments and first validation battery used gpt-4o-mini. We then replicated the validation battery on gpt-5.4-mini using the same prompts, temperature, max-turn limits, and scoring logic, except for a post-hoc exclusion rule for premature NATURAL finalizations described in Section 5. Sampling used a temperature of 0.7.

4.2 Conditions and terminology

We compare four communication formats. To avoid a misleading label, we refer to the compressed protocol by the names used in the paper rather than the raw export labels. The final protocol uses fixed verbs such as FETCH, DATA, PASS, and FUND, and is therefore a compressed fixed-verb protocol rather than a truly numeric language; the raw data files call it NUMERIC and NUMERIC_V2.

Paper label	Raw export label	Description
NATURAL	NATURAL	Plain English, no format restriction
NATURAL-TERSE	NATURAL_TERSE	Plain English, responder limited to one sentence per data point
STRUCTURED	STRUCTURED	JSON messages, for example {"action": "query", "fields": [...], "turn": 1}
COMPRESSED-v1	NUMERIC	Initial, unconstrained compressed verb protocol
COMPRESSED-v2	NUMERIC_V2	Redesigned, fixed-verb compressed protocol

Table 1. Communication conditions and the mapping between paper-facing labels and raw export labels.

The approximate per-message overhead of each format, beyond the information being conveyed, is negligible for natural language because it is carried in the model weights, around fifty tokens for the JSON scaffolding of STRUCTURED, and around one hundred and fifty tokens for the protocol definition of COMPRESSED-v2, paid once in the system prompt.

4.3 Tasks

Five pilot experiments probe different information architectures. Experiment 1 is a binary guessing game with eight options under confirm-or-deny rules. Experiment 2 is a clinical-trial enumeration over eight fields ending in a go or no-go decision. Experiment 3 is a strategic venture-capital triage over twelve fields with three decisive negative signals and an explicit incentive to terminate early. Experiment 4 is a free-form version of the venture task with no procedural constraints, intended to let each format find its natural equilibrium. Experiment 5 re-runs the venture task with the redesigned compressed protocol.

The validation battery comprises four suites and 160 runs in total. Suite A is a scaled known-schema replication at twenty runs per condition. Suite B is a verbosity control introducing NATURAL-TERSE at ten runs per condition. Suite C is an unknown-schema discovery task at ten runs per condition, in which solvers are not given the field list. Suite D is a clinical cross-domain transfer over three scenarios at ten runs per condition, including clear and mixed-evidence decisions.

4.4 Metrics

For each run we record total tokens, summed over input and output across all turns and both agents; the number of turns to completion; accuracy as a boolean on the final decision; format compliance on a zero-to-ten scale measuring surface adherence to the assigned format; task coherence and role coherence on comparable scales; and a count of unsupported requests, meaning queries for information absent from the dataset. We stress that format compliance and task coherence are distinct. In Experiment 1 a 22,165-token deadlock loop received a perfect format-compliance score, because the agents never violated the surface format even though the conversation failed entirely. Compliance measures form; coherence measures whether the conversation accomplished anything. Neither is a substitute for correctness, and we do not use compliance as a quality indicator.

Replication package. The browser harnesses, raw JSON exports, analysis script, and paper artifact are available in the public repository listed in Appendix A.

5. Results

We present results in three acts: the pilot experiments that established task-dependency and identified the failure modes of the early compressed protocol; the redesign that turned that protocol from worst to best; and the validation battery that tested whether the redesign holds at scale and across domains.

5.1 Act 1: Pilot experiments

5.1.1 Experiment 1: binary guessing game

When the task already forces minimal responses, natural language wins. The task here required answers of the form "Is the code X?" and "No", which are close to maximally terse already. NATURAL averaged 1,685 tokens at 100 per cent accuracy. STRUCTURED added fifty-nine per cent overhead for no benefit, because JSON wrapping inflates messages that were already short. COMPRESSED-v1 was unstable: one of five runs collapsed into a twenty-turn deadlock of 22,165

tokens when the agents swapped roles and oscillated between two candidate answers. The lesson is that the efficiency advantage of structured and compressed formats requires verbose natural language to compress; below roughly three to five tokens per natural response, protocol overhead dominates.

Condition	Avg tokens	Avg turns	Accuracy
NATURAL	1,685	5.0	100%
STRUCTURED	2,683	5.0	100%
COMPRESSED-v1	~6,840*	~7.8	100%

Table 2. Experiment 1, binary guessing game ($n=5$ per condition). *The COMPRESSED-v1 mean is heavily distorted by a single 22,165-token deadlock run; its other runs were comparable to STRUCTURED.

5.1.2 Experiment 2: clinical-trial enumeration

When natural language is allowed to be verbose, structured formats win. Here each field came with two to four sentences of clinical interpretation in the natural-language condition. STRUCTURED and COMPRESSED-v1 both improved on NATURAL by roughly thirty-eight to forty per cent and were almost identical to one another, separated by about 229 tokens, or 3.3 per cent. Every condition took exactly nine turns, because the task forced full enumeration of all eight fields. This experiment therefore measures format overhead rather than reasoning strategy, and the saving came almost entirely from replacing the responder's prose with bare values.

Condition	Avg tokens	Avg turns	Accuracy
NATURAL	~11,098	9.0	100%
STRUCTURED	~6,882	9.0	100%
COMPRESSED-v1	~6,653	9.0	100%

Table 3. Experiment 2, clinical-trial enumeration ($n=5$ per condition).

5.1.3 Experiment 3: strategic venture triage

This task introduced twelve fields, three decisive negative signals, and an explicit instruction to stop querying once confident. STRUCTURED won decisively at 4,008 average tokens. COMPRESSED-v1 averaged 9,067 tokens, more than NATURAL at 6,312, and exhibited two distinct pathologies. First, it over-enumerated, querying on average 10.2 fields against 5.4 to 5.8 for the other conditions, which is the Enumeration Problem. Second, in one run it received all three knockout signals and nonetheless recommended funding, which is the Calibration Problem. Inspection of the transcripts identified three specific defects in the compressed protocol: unconstrained verb invention, the absence of any deliberation mechanism, and an ambiguous terminal signal. These defects, rather than the compressed format as such, were the suspected causes.

5.1.4 Experiment 4: free-form

With all procedural constraints removed, STRUCTURED again won at 1,641 average tokens and 100 per cent accuracy, settling naturally into a batch-query pattern. COMPRESSED-v1 collapsed to 6,551 average tokens at sixty per cent accuracy, including one run of 21,093 tokens in which the kickstart message triggered a connection-handshake metaphor and the agents cycled through twenty invented verbs without ever reaching the task. The three defects identified in Experiment 3 were confirmed as the causes.

5.2 Act 2: Protocol redesign (Experiment 5)

Experiment 5 asked a precise question: were the COMPRESSED-v1 failures caused by the compressed format, or by three specific implementation defects? We fixed only those three

defects, adding no benchmark hints and no severity flags, and held the dataset, task, and information density constant. To preserve fairness, the STRUCTURED condition was also restricted to bare values with no optional rationale benchmarks.

Defect in COMPRESSED-v1	Fix in COMPRESSED-v2	Effect
Kickstart "READY 9 agent-A" read as a network handshake	Natural-language kickstart instructing the solver to begin	Eliminated the handshake loop
Unconstrained verb invention (ANALYZE, EVALUATE, ASSESS)	Verbs restricted to FETCH, DATA, PASS, FUND only	Forced internal deliberation; removed extra turns
Terminal "DECIDE 9 FINAL DECISION: FUND" was ambiguous	Terminal decision is the bare word PASS or FUND	Removed terminal ambiguity and the related calibration failure

Table 4. The three protocol defects of COMPRESSED-v1 and their corresponding fixes in COMPRESSED-v2.

The effect of fixing three defects on the same compressed format was dramatic. Token consumption fell by 84.1 per cent, from 6,551 to 1,044; turns fell from 7.0 to 2.0; accuracy rose by forty percentage points, from sixty to one hundred per cent; and the token range collapsed from a fourteen-fold spread to zero.

Condition	Avg tokens	Avg turns	Accuracy	Token range
NATURAL	4,589	3.0	100%	1,843 to 7,833
STRUCTURED	1,481	2.4	100%	1,176 to 2,002
COMPRESSED-v2	1,044	2.0	100%	1,044 to 1,044

Table 5. Experiment 5, redesigned protocol on the venture task ($n=5$ per condition). COMPRESSED-v2 was byte-for-byte identical across all five runs

COMPRESSED-v2 produced an identical conversation on every run. The complete exchange, in all five runs, was the following.

```
[SOLVER, turn 1]
FETCH revenue-growth monthly-burn cash-runway mrr customer-count monthly-churn
team-size tam prior-funding product-stage nps competitor-count

[TASKER, turn 1]
DATA revenue-growth:-38% monthly-burn:$185,000/month cash-runway:10 months
mrr:$41,000 customer-count:14 monthly-churn:31% team-size:9 tam:$3.8B
prior-funding:$450k angel round product-stage:closed beta nps:38
competitor-count:4 well-funded competitors

[SOLVER, turn 2]
PASS
```

Three messages, two turns, 1,044 tokens, the correct answer, and zero variance. The mechanism is that restricting the verb set collapses the space of valid messages. Given knowledge of all twelve field names, the only rational opening move is to fetch them in a single batch; once the data unambiguously determine the decision, the only rational closing move is the correct terminal word. When the possibility space contains a single rational path, every run takes it. This reframes the Calibration Problem identified earlier: it was a symptom of protocol ambiguity dividing the model's effort between interpreting the protocol and reasoning about the task, rather than an intrinsic property of compressed values.

5.3 Act 3: Validation battery

The battery ran all four suites to completion with no failed API runs. Aggregating the comparable conditions across all suites, that is excluding the NATURAL-TERSE control, gives the headline picture.

Condition	Runs	Accuracy	Avg tokens	Avg turns
NATURAL	50	48/50 (96%)	6,259	3.88
STRUCTURED	50	49/50 (98%)	1,030	2.06
COMPRESSED-v2	50	48/50 (96%)	847	2.12

Table 6. Aggregate cross-suite results over comparable conditions (50 runs per condition).

Three points stand out. STRUCTURED is the accuracy leader overall. COMPRESSED-v2 is the token-efficiency leader, but its advantage over STRUCTURED is only about 183 tokens per run, whereas the advantage of leaving natural language is more than five thousand tokens per run. The large efficiency gain therefore comes from abandoning natural language, not from replacing JSON with a compressed protocol.

5.3.1 Suite A: scaled known-schema replication

Suite A confirmed the Experiment 5 result at twenty runs per condition. All three conditions achieved perfect accuracy, but the cost and variance differed sharply. COMPRESSED-v2 averaged 772 tokens within a range of 767 to 775, around twenty-six per cent cheaper than STRUCTURED. NATURAL remained accurate but extremely variable, reaching 34,112 tokens on one run, roughly forty-four times a typical COMPRESSED-v2 run. The effect sizes against NATURAL were large, with Cohen's d of minus 1.54 for STRUCTURED and minus 1.60 for COMPRESSED-v2.

Condition	n	Accuracy	Avg tokens	Range	Avg turns
NATURAL	20	20/20	8,660	1,724 to 34,112	4.35
STRUCTURED	20	20/20	1,045	966 to 1,075	2.00
COMPRESSED-v2	20	20/20	772	767 to 775	2.00

Table 7. Suite A, scaled known-schema venture task (n=20 per condition).

This result is partly a property of the task: when all fields are known, the rational move is to fetch every field at once, which is exactly where a compressed protocol should excel. Suite A therefore establishes a strong upper bound on the compressed advantage, which the remaining suites qualify.

5.3.2 Suite B: verbosity control

Suite B tested whether NATURAL is expensive because it is natural language or because it is verbose, by constraining the responder to one sentence per data point. The answer is both, with verbosity the larger component. NATURAL-TERSE cut token use by about fifty-seven per cent relative to NATURAL, a substantial reduction, yet it still used about 2.6 times as many tokens as STRUCTURED and about 3.5 times as many as COMPRESSED-v2, and it took more turns on average. Crucially, its task-coherence score fell to 6.5 out of ten.

Condition	n	Accuracy	Avg tokens	Reduction vs NATURAL	Coherence
NATURAL	10	10/10	6,259	baseline	9.8
NATURAL-TERSE	10	10/10	2,699	-56.9%	6.5
STRUCTURED	10	10/10	1,057	-83.1%	10.0
COMPRESSED-v2	10	10/10	771	-87.7%	10.0

Table 8. Suite B, NATURAL-TERSE verbosity control (n=10 per condition).

The coherence drop reveals why terseness is insufficient. NATURAL-TERSE solvers continued to ask broad due-diligence questions about founding teams, product differentiation, projected financials, and intellectual property, none of which existed in the dataset, and the responder repeatedly had to decline. Terse natural language reduces output length but does not convert a natural-language agent into a database-querying agent. The format still activates a broad real-world schema that verbosity constraints do not suppress.

5.3.3 Suite C: unknown-schema discovery

Suite C removed the field list from all solvers, so that the schema had to be discovered rather than batched. Here the efficiency gap between STRUCTURED and COMPRESSED-v2 nearly disappeared, at 1,126 against 1,162 average tokens. NATURAL remained expensive at 3,322 tokens but was perfectly accurate, because its broad questioning sometimes surfaced the decisive fields. STRUCTURED suffered its single accuracy failure of the battery: in one run the solver discovered only favourable fields, namely prior funding, total addressable market, and net promoter score, missed the decisive negatives, and recommended funding.

Condition	n	Accuracy	Avg tokens	Median	Avg turns
NATURAL	10	10/10	3,322	3,548	3.30
STRUCTURED	10	9/10	1,126	907	2.30
COMPRESSED-v2	10	10/10	1,162	805	2.60

Table 9. Suite C, unknown-schema discovery (n=10 per condition).

Suite C prevents over-claiming from Suite A. In unknown-schema settings, the advantage of compression narrows, and natural language's broad schema can elicit relevant information that a discovery-limited structured query may miss. Notably, COMPRESSED-v2 remained perfectly accurate here, which suggests its category-token affordances partially solved the discovery problem without a wildcard fetch-all command.

5.3.4 Suite D: clinical cross-domain transfer

Suite D rotated three clinical scenarios, a clear approval, a clear rejection, and a mixed-evidence rejection, to test whether the findings transfer outside the venture domain. In the original gpt-4o-mini battery, STRUCTURED was the standout, achieving ten out of ten at 877 average tokens. COMPRESSED-v2 remained extraordinarily cheap at 755 tokens within a range of 754 to 758, but failed twice, both times on the mixed-evidence scenario. NATURAL also failed twice, but by a

different mechanism: it demanded secondary endpoints and subgroup analyses that did not exist, became over-cautious, and rejected a drug that the data supported approving.

Condition	n	Accuracy	Avg tokens	Range	Avg turns
NATURAL	10	8/10	4,397	2,626 to 7,213	3.60
STRUCTURED	10	10/10	877	833 to 904	2.00
COMPRESSED-v2	10	8/10	755	754 to 758	2.00

Table 10. Suite D, clinical cross-domain transfer (n=10 per condition).

Scenario	NATURAL	STRUCTURED	COMPRESSED-v2
DRUG-7X, clear approval	3/4	4/4	4/4
DRUG-9Q, clear rejection	3/3	3/3	3/3
DRUG-4M, mixed rejection	2/3	3/3	1/3

Table 11. Suite D accuracy by scenario in the original gpt-4o-mini battery. The mixed-evidence case is where COMPRESSED-v2 fails.

In the failing original COMPRESSED-v2 runs the protocol surfaced every relevant field correctly, including a moderate cardiovascular safety signal, low trial completion, mixed regulatory precedent with black-box warnings, and an unfunded post-market registry, and the solver still chose to proceed. The transcript records what the solver saw but not how it weighed the evidence, because the format leaves no room for deliberation. STRUCTURED, retaining a minimal rationale field, weighed the same mixed evidence correctly. We stress that the labels in Suite D are synthetic task labels defined by the dataset design and are not medical recommendations or expert clinical judgements.

Cross-model replication on gpt-5.4-mini

We repeated the full validation battery on gpt-5.4-mini. The run produced all 160 expected trials with no API failures. Eight NATURAL runs emitted a final decision before receiving a Tasker response; seven of those premature finalizations happened to match the expected label, but they are not valid agent-to-agent exchanges. The adjusted aggregate below excludes those eight one-message NATURAL runs and leaves the structured and compressed conditions unchanged.

Condition	Runs	Accuracy	Avg tokens	Avg turns
NATURAL	42	41/42 (97.6%)	2,215	2.00
STRUCTURED	50	50/50 (100%)	1,209	2.12
COMPRESSED-v2	50	50/50 (100%)	788	2.00

Table 12. GPT-5.4-mini validation replication, adjusted aggregate over comparable conditions. Eight premature NATURAL one-message finalizations are excluded from the NATURAL aggregate.

The replication strengthens the central efficiency result while changing the reliability interpretation. COMPRESSED-v2 achieved fifty out of fifty on gpt-5.4-mini at 788 average tokens, including ten out of ten in the clinical transfer suite and three out of three on the mixed-evidence clinical case. The original gpt-4o-mini mixed-evidence failures should therefore be read as model-dependent reliability failures rather than as proof that compressed protocols intrinsically cannot weigh mixed evidence. The interpretability cost remains, because the successful compressed runs still expose little about how evidence was weighted.

6. Discussion

6.1 What agents gain

The efficiency gains are real and large. Across information-transfer tasks, structured and compressed formats reduced token use substantially relative to natural language. They also delivered far greater determinism. In Suite A, NATURAL ranged from 1,724 to 34,112 tokens, whereas COMPRESSED-v2 ranged only from 767 to 775 in the original battery and remained near 790 tokens in the gpt-5.4-mini replication. This has direct operational significance, because a system with a predictable cost ceiling is categorically easier to operate than one with a fat-tailed cost distribution. Structured formats additionally enforce information boundaries automatically through syntax, which is the Information Containment Problem turned to advantage, and is a genuine property difference rather than experimental noise.

6.2 Verbosity is separable from natural language

Suite B shows that verbosity is a large component of natural language's cost, since constraining it cut tokens by fifty-seven per cent, but it is not the whole story, because terse natural language still cost between 2.5 and 3.5 times as much as structured formats. The residual gap is the broad-schema activation effect: natural-language agents model what information would matter in the real world, not only what is available in the dataset, and they ask about things that do not exist. Format constraints suppress this behaviour; linguistic constraints alone do not.

6.3 Protocol design quality is the dominant variable

The comparison between the two versions of the compressed protocol is the clearest result in the study. Three specific design changes on the same format produced an eighty-four per cent reduction in tokens, a forty-percentage-point increase in accuracy, and a collapse in variance from a fourteen-fold range to zero. The format category was constant; only the protocol design changed. The practical question for designers of multi-agent systems is therefore not which format to use, but how well the protocol is designed within a format. A poorly designed compressed protocol is worse than natural language; a well-designed one beats JSON on both cost and determinism.

6.4 What agents lose: reliability

Compressed formats expose two predictable risk zones. The first is unknown-schema discovery, where the solver must guess what to ask for and the efficiency advantage of compression narrows; in our original battery, the observed discovery failure was STRUCTURED rather than COMPRESSED-v2, but the result still shows that schema discovery changes the communication problem. The second is mixed-signal evidence, where no single value dominates and the decision requires weighing competing factors. In the original gpt-4o-mini Suite D clinical case, COMPRESSED-v2 surfaced all the evidence correctly but overweighted statistically significant efficacy and underweighted safety, completion, and adverse regulatory precedent. In the gpt-5.4-mini replication, that compressed-format clinical failure disappeared: COMPRESSED-v2 achieved ten out of ten in Suite D, including three out of three on the mixed-evidence case. The reliability cost is therefore model-dependent rather than intrinsic to the compressed protocol, but the auditability cost remains.

6.5 What agents lose: interpretability

When COMPRESSED-v2 fails, the transcript is three lines with no reasoning trace, and there is no way to audit why the decision was wrong. Even when COMPRESSED-v2 does not fail, the transcript gives much less evidence about how the solver weighted competing factors. A STRUCTURED failure includes at least a rationale sentence, and a NATURAL failure includes full reasoning that can be examined. This is not a speculative concern; it is demonstrated directly in the original Suite D failure cases. As compressed machine-to-machine communication becomes more capable, the risk is that efficiency gains crowd out the interpretability that allows humans to catch and diagnose errors. The format that makes the channel efficient is the same format that makes the reasoning opaque, which connects this work to the argument that externalised reasoning is a fragile but valuable oversight opportunity [14].

6.6 The English-Brain Problem as the unifying explanation

The efficiency gains are real because the channel is genuinely wasteful: natural language often generates tokens that carry little additional information beyond the underlying data. The failure modes are consistent with the English-Brain framework because current models have acquired much of their domain knowledge through human linguistic framing, and stripping that framing from incoming data can remove cues that help the relevant concepts activate. One can make the mouth smaller, but the brain stays the same size. This is not a limitation of any single protocol but a structural hypothesis about current architectures trained on human text. Escaping it would require either fine-tuning a protocol into the model weights or training on fundamentally different data.

6.7 STRUCTURED as the robust middle ground

STRUCTURED was the overall accuracy leader in the original battery at forty-nine of fifty and achieved fifty of fifty in the gpt-5.4-mini replication. It never produced a catastrophic loop or near-empty audit trail, although it did have one discovery failure in Suite C of the original battery. It preserves semantic labels that let domain knowledge activate correctly, enforces information boundaries through syntax, and produces minimal but legible reasoning traces. For production multi-agent systems in which decisions carry real consequences, the ratio of robustness to cost makes STRUCTURED the conservative option in this study. The compressed protocol remains the right choice where the schema is known, the evidence is decisive, predictable minimal cost matters most, and the surrounding system can tolerate reduced auditability.

7. Limitations

- The experiments used two OpenAI models, gpt-4o-mini and gpt-5.4-mini. The replication increases confidence in the efficiency result, but additional model families are needed before strong general claims are made.
- The tasks are synthetic with hand-designed schemas. Real deployment tasks have messier information architectures, and the clean batch-and-decide optimum of Suite A may not generalise.
- Sample sizes in Suites B, C, and D are small at ten runs per condition in each battery, so those results remain directional. The cross-model replication increases confidence but does not replace larger-n studies.

- The experiments were run in a browser harness rather than a packaged, reproducible script. A separate analysis script that reads the exported JSON and regenerates all tables should accompany a final submission. Because token counts are skewed and NATURAL has fat-tailed runs, final statistical reporting should emphasise medians, bootstrap confidence intervals, and non-parametric or permutation tests in addition to means and Cohen's d .
- The format-compliance metric is unreliable as a quality indicator, as the Experiment 1 deadlock demonstrates. Task and role coherence address this but were added mid-study. None of these should be treated as a substitute for correctness.
- Both validation-battery runs used a temperature of 0.7. Results may differ under deterministic decoding or other sampling settings.
- Suite D correct answers are defined by dataset design rather than by expert clinical judgement, and no clinician validated them. They are not medical recommendations.
- The bootstrapping cost was not measured. Every condition received its protocol description in natural language in the system prompt, and the one-time cost of specifying a compressed protocol was not included in the token counts.

8. Future Work

- Replicate Suites A to D on additional model families, including non-OpenAI frontier models, to test whether the protocol trade-offs generalise.
- Scale Suites C and D to twenty or thirty runs per condition, since these are where the interesting robustness trade-offs appear.
- Fine-tune a model on the compressed protocol to test whether the bootstrapping cost disappears and whether calibration improves.
- Add a fifth condition using native function-calling or tool APIs, which are structured but supported directly by the model.
- Test fully hidden schemas with no category affordances in the discovery suite, to locate the boundary of compressed-format discovery.
- Add human expert evaluation of the clinical decisions in Suite D.
- Conduct an economic threshold analysis: at what volume of machine-to-machine communication does the per-conversation saving become operationally meaningful relative to the reliability risk?

9. Conclusion

GibberLink posed a question in the voice layer: if two AI systems are talking and no human is listening, why are they speaking a human language? We ran the text-domain equivalent of that question systematically, comparing natural language, terse natural language, JSON-structured messages, and a compressed fixed-verb protocol across five pilot experiments, a 160-run validation battery, and a 160-run gpt-5.4-mini replication.

The answer is nuanced. Natural language is expensive and high in variance. A well-designed compressed protocol is efficient and deterministic in bounded, known-schema tasks. JSON-structured communication is the most robust conservative format overall, while the gpt-5.4-mini replication shows that stronger models can preserve compressed-protocol reliability even in the clinical mixed-evidence suite. The efficiency gains of leaving natural language are real, but they come with task-dependent reliability risks and a persistent interpretability cost: compressed formats lose much of their advantage in discovery settings and leave little auditable trace when they err.

The central lesson is that the optimal format for agent-to-agent communication is not natural language, JSON, or compressed tokens universally. It depends on the information architecture of the task, and within any format, the quality of the protocol design matters more than the format category itself. The channel between agents can be compressed; the English-shaped cognition inside each agent cannot, and that asymmetry explains both what agents gain and what they lose when they stop speaking English.

References

- [1] B. Starkov and A. Pidkuiko, "GibberLink: two conversational AI agents switching from English to a sound-level protocol after confirming both are AI," software demonstration, ElevenLabs Worldwide Hackathon, 2025. Uses the ggwave data-over-sound library by G. Gerganov. Source code: <https://github.com/PennyroyalTea/gibberlink> (accessed 20 June 2026).
- [2] Model Context Protocol project, "What is the Model Context Protocol (MCP)?" open protocol documentation, 2024-2026. Available: <https://modelcontextprotocol.io/> (accessed 20 June 2026).
- [3] A2A Protocol project, "Agent2Agent (A2A) Protocol," open agent-interopability specification, originally developed by Google and donated to the Linux Foundation, 2025-2026. Available: <https://a2a-protocol.org/latest/> (accessed 20 June 2026).
- [4] Y. Labrou and T. Finin, "Semantics and conversations for an agent communication language," in Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan, 1997. arXiv: cs/9809034.
- [5] Foundation for Intelligent Physical Agents, "FIPA ACL Message Structure Specification," standard SC00061, 2002.
- [6] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," arXiv:2006.02419, 2020.
- [7] A. Lazaridou, A. Potapenko, and O. Tieleman, "Multi-agent communication meets natural language: synergies between functional and structural language learning," arXiv:2005.07064, 2020.
- [8] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: communicative agents for mind exploration of large language model society," arXiv:2303.17760, 2023.
- [9] Q. Wu, G. Bansal, J. Zhang, Y. Wu, et al., "AutoGen: enabling next-generation LLM applications via multi-agent conversation," arXiv:2308.08155, 2023.
- [10] S. Hong, X. Zheng, J. Chen, Y. Cheng, et al., "MetaGPT: meta programming for a multi-agent collaborative framework," arXiv:2308.00352, 2023.
- [11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, et al., "Chain-of-thought prompting elicits reasoning in large language models," arXiv:2201.11903, 2022.
- [12] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting," arXiv:2305.04388, 2023.
- [13] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, et al., "Measuring faithfulness in chain-of-thought reasoning," arXiv:2307.13702, 2023.
- [14] T. Korbak, M. Balesni, E. Barnes, et al., "Chain of thought monitorability: a new and fragile opportunity for AI safety," arXiv:2507.11473, 2025.
- [15] M. A. Hakim, "Brevity Constraints Reverse Performance Hierarchies in Language Models," arXiv:2604.00025, 2026.
- [16] J. Brussee, "Caveman: a token-compression skill for coding assistants," software, 2025. Source code: <https://github.com/juliusbrussee/caveman> (accessed 20 June 2026).
- [17] P. Singhal, T. Goyal, J. Xu, and G. Durrett, "A long way to go: investigating length correlations in RLHF," arXiv:2310.03716, 2023.
- [18] W. Shen, R. Zheng, W. Zhan, J. Zhao, et al., "Loose lips sink ships: mitigating length bias in reinforcement learning from human feedback," arXiv:2310.05199, 2023.

- [19] E. Zelikman, G. Harik, Y. Shao, V. Jayasiri, N. Haber, and N. D. Goodman, “Quiet-STaR: language models can teach themselves to think before speaking,” arXiv:2403.09629, 2024.
- [20] S. Hao, S. Sukhbaatar, D. Su, X. Li, et al., “Training large language models to reason in a continuous latent space (Coconut),” arXiv:2412.06769, 2024.
- [21] X. Zou, J. Huang, J. Li, and P. Zhou, “Latent Thought Flow: Efficient Latent Reasoning in Large Language Models,” arXiv:2606.16222, 2026. Cited cautiously as very recent work.
- [22] Y. Zhang, B. Tang, T. Ju, S. Duan, and G. Liu, “Do Latent Tokens Think? A Causal and Adversarial Analysis of Chain-of-Continuous-Thought,” arXiv:2512.21711, 2025.
- [23] W. Chen, C. Yuan, J. Yuan, Y. Su, C. Qian, C. Yang, R. Xie, Z. Liu, and M. Sun, ‘Beyond Natural Language: LLMs Leveraging Alternative Formats for Enhanced Reasoning and Communication,’ in Findings of the Association for Computational Linguistics: EMNLP 2024, arXiv:2402.18439, 2024.

Appendix A: Experiment and Data Mapping

The pilot experiments and original validation battery used gpt-4o-mini at a temperature of 0.7. The validation battery was also replicated on gpt-5.4-mini using the same temperature and scoring logic, with premature NATURAL one-message finalizations excluded from the adjusted aggregate. The replication package is available in the public repository below.

<https://github.com/Altyd/machine-native-agent-communication-replication>

Experiment / Suite	Task	Conditions	n	Data file
Experiment 1	Binary guessing game, 8 options	NAT, STR, C-v1	5	data/pilot_01_binary_guessing_game.json
Experiment 2	Clinical enumeration, 8 fields	NAT, STR, C-v1	5	data/pilot_02_clinical_trial.json
Experiment 3	Venture triage, 12 fields	NAT, STR, C-v1	5	data/pilot_03_strategic_triage.json
Experiment 4	Free-form venture decision	NAT, STR, C-v1	5	data/pilot_04_freeform_vc.json
Experiment 5	Venture task, redesigned protocol	NAT, STR, C-v2	5	data/pilot_05_compressed_v2.json
Suite A	Scaled known-schema venture	NAT, STR, C-v2	20	data/battery_gpt40_mini.json; data/battery_gpt54_mini.json
Suite B	Verbosity control	NAT, NAT-TERSE, STR, C-v2	10	data/battery_gpt40_mini.json; data/battery_gpt54_mini.json
Suite C	Unknown-schema discovery	NAT, STR, C-v2	10	data/battery_gpt40_mini.json; data/battery_gpt54_mini.json
Suite D	Clinical cross-domain, 3 scenarios	NAT, STR, C-v2	10	data/battery_gpt40_mini.json; data/battery_gpt54_mini.json

Table A1. NAT = NATURAL, NAT-TERSE = NATURAL-TERSE, STR = STRUCTURED, C-v1 = COMPRESSED-v1, C-v2 = COMPRESSED-v2.

Appendix B: Full Per-Suite Statistics

The following tables reproduce the detailed statistics for each validation suite, including 95 per cent confidence intervals, medians, and format and coherence scores. Effect sizes are Cohen's d relative to NATURAL.

Suite A (n=20 per condition)

Condition	Acc.	Avg tok.	95% CI	Median	Range	Turns	d vs NAT
NATURAL	20/20	8,660	$\pm 3,057$	7,106	1,724-34,112	4.35	-
STRUCTURED	20/20	1,045	± 18	1,065	966-1,075	2.00	-1.54
COMPRESSED-v2	20/20	772	± 2	775	767-775	2.00	-1.60

Table B1.

Suite B (n=10 per condition)

Condition	Acc.	Avg tok.	95% CI	Median	Range	Coher.
NATURAL	10/10	6,259	$\pm 1,260$	6,515	1,654-8,347	9.8
NATURAL-TERSE	10/10	2,699	± 763	2,410	1,002-4,709	6.5
STRUCTURED	10/10	1,057	± 21	1,069	962-1,073	10.0
COMPRESSED-v2	10/10	771	± 3	771	767-775	10.0

Table B2.

Suite C (n=10 per condition)

Condition	Acc.	Avg tok.	95% CI	Median	Range	Turns
NATURAL	10/10	3,322	± 973	3,548	1,270-6,083	3.30
STRUCTURED	9/10	1,126	± 434	907	881-3,117	2.30
COMPRESSED-v2	10/10	1,162	± 361	805	794-2,036	2.60

Table B3.

Suite D (n=10 per condition)

Condition	Acc.	Avg tok.	95% CI	Median	Range	Turns
NATURAL	8/10	4,397	± 939	4,445	2,626-7,213	3.60
STRUCTURED	10/10	877	± 13	881	833-904	2.00
COMPRESSED-v2	8/10	755	± 1	754	754-758	2.00

Table B4.

GPT-5.4-mini adjusted aggregate

The replication battery produced all 160 expected trials with no API failures. Eight NATURAL runs emitted a final decision before receiving any Tasker response; these are labeled as premature finalization / invalid communication runs and excluded from the adjusted NATURAL aggregate.

Condition	Runs	Accuracy	Avg tokens	Avg turns
NATURAL	42	41/42 (97.6%)	2,215	2.00
NATURAL-TERSE	10	10/10 (100%)	1,407	2.00
STRUCTURED	50	50/50 (100%)	1,209	2.12
COMPRESSED-v2	50	50/50 (100%)	788	2.00

Table B5. GPT-5.4-mini replication, adjusted aggregate. NATURAL excludes eight premature one-message finalizations; the other conditions are unchanged.

GPT-5.4-mini Suite D by scenario

Scenario	NATURAL	STRUCTURED	COMPRESSED-v2
DRUG-7X, clear approval	2/3	4/4	4/4
DRUG-9Q, clear rejection	2/2	3/3	3/3
DRUG-4M, mixed rejection	1/1	3/3	3/3

Table B6. GPT-5.4-mini Suite D accuracy by scenario after excluding premature NATURAL finalizations.

Excluded premature NATURAL runs

Suite	Run	Scenario	Expected	Correct
A	8	vc-nexus-pass	PASS	True
C	1	vc-nexus-pass	PASS	True
C	2	vc-nexus-pass	PASS	True
C	8	vc-nexus-pass	PASS	True
D	3	drug-4m-mixed	NO-GO	True
D	5	drug-9q-nogo	NO-GO	True
D	9	drug-4m-mixed	NO-GO	True
D	10	drug-7x-go	GO	False

Table B7. Premature NATURAL one-message finalizations excluded from the adjusted GPT-5.4-mini aggregate.

Appendix C: Representative Transcripts

C.1 COMPRESSED-v2 successful bounded run (Experiment 5, Suite A)

The complete and identical conversation produced on every run of the redesigned protocol in the bounded known-schema venture task.

```

FETCH revenue-growth monthly-burn cash-runway mrr customer-count monthly-churn
      team-size tam prior-funding product-stage nps competitor-count
DATA revenue-growth:-38% monthly-burn:$185,000/month cash-runway:10 months
      mrr:$41,000 customer-count:14 monthly-churn:31% team-size:9 tam:$3.8B
      prior-funding:$450k product-stage:closed beta nps:38 competitor-count:4
PASS

```

C.2 COMPRESSED-v1 handshake loop (Experiment 4)

The most catastrophic early failure. A kickstart message read as a network handshake caused the agents to cycle through invented connection verbs for twenty turns and 21,093 tokens without ever reaching the task.

```

READY 9 agent-A

```

```
INITIATE -> CONNECT -> TRANSMIT -> DELIVER -> EXECUTE -> FINALIZE ->
COMPLETE -> TERMINATE -> CLOSE -> DISCONNECT -> ABORT -> HALT -> STOP ->
RELEASE -> WITHDRAW -> SUSPEND -> PAUSE -> CEASE -> FINISH -> CONCLUDE
(20 turns, 21,093 tokens, no decision reached)
```

C.3 STRUCTURED discovery miss (Suite C)

The single STRUCTURED accuracy failure in the battery. The solver discovered only favourable fields and never queried the decisive negatives, then recommended funding. Expected: PASS. Actual: FUND.

```
{"action":"query","fields":["funding","market","team","product"],"turn":1}
{"action":"respond","fields":[{"prior-funding":"$450k"}, {"tam":"$3.8B"},
 {"team-size":"9"}, {"product-stage":"closed beta"}, {"nps":"38"}], "turn":1}
{"action":"decide","decision":"FUND","rationale":"Solid prior funding, a large
total addressable market, and a positive NPS indicate strong growth potential.",
"turn":2}
```

The decisive negatives that were never discovered: revenue growth minus thirty-eight per cent, monthly churn thirty-one per cent, cash runway ten months, monthly burn \$185,000.

C.4 Original gpt-4o-mini COMPRESSED-v2 mixed clinical miss (Suite D, DRUG-4M)

An original gpt-4o-mini compressed-format failure on mixed evidence. Every relevant field was surfaced, yet the solver chose to proceed, leaving no record of how it weighed the evidence. Expected: NO-GO. Actual: GO.

```
FETCH efficacy-treatment efficacy-placebo adverse-events trial-completion
p-value safety-signal regulatory-precedent post-market-commitment
DATA efficacy-treatment:63.8% efficacy-placebo:41.2% adverse-events:29.5%
trial-completion:72.0% p-value:0.018 safety-signal:Moderate cardiovascular
imbalance regulatory-precedent:Mixed; two approvals required black-box
warnings post-market-commitment:36-month registry proposed but not funded
GO
```

C.5 NATURAL clinical over-caution (Suite D, DRUG-7X)

A natural-language failure by the opposite mechanism. The solver demanded secondary endpoints, subgroup analyses, and regulatory feedback that the dataset did not contain, and rejected a drug the data supported approving. Expected: GO. Actual: NO-GO. This illustrates that the broad schema natural language activates can be a liability in bounded datasets, just as it can be an asset in open-ended ones.

End of document