UDC 004.8:621.391(045) DOI 10.18372/2073-4751.87.21694

A. Arutiunian[1],
orcid.org/0009-0004-9785-4762,
e-mail: arto@urobora.com,
O. Tachinina[1,2],
orcid.org/0000-0001-7081-0576,
e-mail: olena.tachynina@npp.kai.edu.ua

METHOD FOR DETERMINING THE MAXIMUM SEMANTIC LOAD ON A SHARED CONTEXT FOR A REQUIRED ANSWER ACCURACY

[1]National University «Kyiv Aviation Institute»
[2]National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

Introduction

Multi-agent systems built on large language models (LLMs) have moved from research prototypes to production infrastructure. The design heuristic behind that expansion is simple: sample several reasoning paths, aggregate them by majority vote or learned verification, and expect accuracy to rise with agent count. It draws support from self-consistency decoding [1], which reported monotone gains from sampling additional chain-of-thought paths, and from later work in which ensembles of dozens of homogeneous agents bring weaker models up to the level of stronger ones [2].

Recent results contradict that picture from several directions. For compound systems, accuracy can be non-monotone in the number of LLM calls, with the net effect depending on the task's difficulty distribution [3]. A study of 260 configurations, five architectures and three model families observed error amplification of up to 17.2× in independent-agent systems and a saturation threshold beyond which coordination yields negative returns [4]. An audit of seven open-source multi-agent frameworks found failure rates of 41–86.7%, with nearly a quarter of all failures attributable to the verification stage rather than to the generating agents [5]. An information-theoretic account locates the benefit of scaling in the number of non-redundant information sources rather than in raw agent count, which saturates early under homogeneous ensembles [6].

Underneath these results sits a constraint usually treated separately: the agents share one context window, and what they can use of that window falls well short of its nominal token limit. Performance degrades as input length grows even where retrieval accuracy is perfect [7], degradation is position-dependent [8], and a maximum effective window can be located empirically well below the advertised limit [9].

These are measurements of a degradation curve rather than of a capacity: none defines a quantity in bits that a converse could bound, so none can say what an orchestration built on the window can and cannot achieve. The closest prior construction, a Fano-style accuracy bound for single-pass multi-hop reasoning [10], bounds accuracy from an assumed information content rather than defining and estimating the capacity itself. Usable information relative to a computationally constrained decoder family [11] gives such a capacity its footing; to our knowledge, that family has not been instantiated as a single call to a fixed reader model with an explicit read budget, and no converse for multi-agent orchestration has been derived from it. Separately, resampling against an imperfect verifier hits a ceiling set by the verifier's precision [12], and LLMs are poor self-verifiers [13]. That literature treats verifier quality as the binding constraint and leaves aside the case in which generator and verifier read the same context and saturate together. Detectors such as semantic entropy

[14] flag unreliable generations post hoc rather than bounding what a shared context can support.

Problem Statement

No definition of shared-context capacity in the work above is simultaneously (a) finite and non-trivial for a *lossless* workspace, (b) independent of which orchestration reads it, and (c) estimable from black-box API access. Without such a quantity, the practical design question, at what load adding agents stops helping, cannot be posed formally.

That question is a design problem of the ordinary engineering kind: a system parameter must be chosen to meet a required performance index. The parameter is how much material may be placed in the shared context; the performance index is the probability of a correct answer; and the difficulty is that no existing theory says where the admissible region ends.

The study concerns multi-agent information-processing systems built on LLMs whose agents read one shared context, and specifically the information capacity of that context at the read interface and its relation to the maximum admissible operating load of the system. The aim of this work is to develop a method that determines, from black-box access to the model alone, the maximum admissible semantic load on the context shared by the agents, with a result that, at a fixed read budget and protocol class, does not depend on the orchestration topology reading the context. A load is admissible if the usable capacity of the context (defined below) does not rule it out. Admissibility is therefore a necessary condition for a specified probability of a correct answer and never shows that the target is reached. The converse below guarantees only the opposite direction: at loads beyond those the capacity of the context supports, no orchestration topology reaches the specified probability.

The inputs are a reader model M accessible only through its API; a query battery in which each query admits exactly v equiprobable answers, so the per-query entropy $h_0 = \log_2 v$ is exact rather than assumed; a target mean per-query error ε_0; a read budget m and a class $\mathscr{B}$ of permitted read protocols; and, where candidates carry a verifiable certificate, a target coverage f_0. From these the method must find two admissible loads, both measured in atomic facts placed in the shared context. The first, $l_{max}^{(\varepsilon)}$, is the greatest load at which the usable capacity of the context does not preclude a mean error of ε_0 or below. The second, $l_{max}^{(f)}$, is the greatest load at which it does not preclude coverage of at least a fraction f_0 of candidates under a certificate-gated policy. The second load gives the price of detecting exhaustion: a gated system signals that it has exhausted its capacity by declining to answer, so detection and abstention coincide, and detecting exhaustion costs the coverage lost to abstention.

Three constraints bind the solution. *Topology invariance:* at a fixed read budget and protocol class, the bound must hold for every orchestration over the shared context, from a single agent to majority vote, centralized verification, debate, chain and graph schemes. *Observability:* every quantity entering the computation must be measurable from black-box access. *One-sidedness:* the capacity estimate must be a lower confidence bound, so that both resulting loads are themselves lower bounds; the reported boundary then lies at or below the true one, and the output can serve as a design parameter.

The paper makes three contributions. First, we define a usable capacity of a shared context at the *read interface*, which makes the quantity finite and non-trivial for a workspace that stores information without loss. Second, we derive a Fano-type converse for this capacity whose envelope, at a fixed read budget and protocol class, does not depend on the orchestration topology. Third, we give a procedure that obtains a lower confidence bound on this capacity from black-box access alone and inverts it into the two admissible loads.

Problem Solution

Capacity at the read interface. We model a multi-agent workflow as a cascade of two channels. A task answer X is rendered into a shared *workspace* W, the context window the agents read, and each agent call maps W to an output. We write ℓ for the *semantic load*, the number of atomic facts placed in the workspace, as distinct from its token count; L denotes the workspace length in tokens and Σ the reader's token vocabulary.

In the considered setting, the first stage is verbatim concatenation at token counts far below the window limit, so it is lossless: the workspace string retains full Shannon information about the stored answers at every load tested. A capacity defined on that string is therefore vacuous. The degradation arises in the second stage, in what the reader can retrieve from a window that probably still contains the answer, and capacity has to be defined where that loss occurs.

For a query battery Q with gold answers X in alphabets of size at most v, an m-read protocol against (W, Q) produces a sequence of outputs $Y_1, \ldots, Y_m$ in which each prompt is an arbitrary function of W, Q and the earlier outputs. Adaptivity is deliberate, so the definition covers parallel ensembles, verifier calls, debate rounds and chains alike. The Markov chain $X \to (W, Q) \to Y_{1:m}$ then holds by construction, provided each call sees the answers only through W.

Definition 1 (Usable workspace capacity; cf. [11]). *For a read budget $m \geq 1$ and a class $\mathscr{B}$ of permitted read protocols,*

$$C_{ws}(m; B) = \sup I(X; Y_{1:m} \mid Q), \quad (1)$$

the supremum over load, query and answer distributions with finite alphabets, encodings into at most L tokens, and m-read protocols drawn from $\mathscr{B}$.

The two arguments do different work: m is how many reads an orchestration spends; $\mathscr{B}$ is what those reads may be. By the data-processing inequality every $C_{\mathrm{ws}}(m; \mathscr{B})$ is at most $L \log_2|\Sigma|$, the *nominal* capacity of the window. This work is organized around the hypothesis that the two are far apart for the class $\mathscr{B}_{\mathrm{task}}$ of *content-addressed* protocols, whose reads must locate an answer inside the workspace rather than transcribe a span named in the prompt. The restriction matters because a span-addressed transcription protocol attains $H(X)$: over an unrestricted class the supremum reaches the nominal ceiling and the converse below is vacuous. No experiment here bounds capacity over the unrestricted class.

The converse. Let $\varphi_v(\varepsilon) = h(\varepsilon) + \varepsilon \log_2(v-1)$ be the v-ary Fano function, with h the binary entropy. The result rests on two assumptions. **B1**: conditional on Q, the answers are mutually independent with $H(X_k \mid Q) \geq h_0 > 0$ bits. **B2** (grounding): the reader has no parametric side channel to the specific answers.

Proposition 1 (Capacity converse for grounded workflows). *Under B1–B2, let a workflow whose reads belong to $\mathscr{B}_{task}$ use at most m reads and produce final answers with mean error $\bar{\varepsilon}(\ell)$. Then*

$$\varphi_v\left(\underline{\varepsilon}(l)\right) \geq h_0 - \frac{C_{ws}(m; B_{task})}{l}, \quad (2)$$

$$C_{ws}(m; B_{task}) \geq l\left(h_0 - \varphi_v\left(\underline{\varepsilon}(l)\right)\right) \quad (3)$$

The argument is elementary given the definition: the chain rule under B1, the fact that the workflow's operating point lies within the supremum in (1), Fano's inequality for each query, and Jensen's inequality, since φ_v is concave. Inequality (2) is an impossibility statement delimiting the admissible operating region of every orchestration over the workspace. Inequality (3) turns an observed error into a lower bound on capacity, and so serves as an estimator.

At a fixed read budget and protocol class, the envelope applies unchanged, with the same C_{ws}, to a single-agent baseline, to independent-parallel majority vote, to centralized verification whether or not the verifier sees the workspace, and to any debate, chain or graph orchestration over the shared workspace satisfying B2. Hence, for orchestrations whose reads locate answers in

the workspace, no aggregation scheme or topology redesign restores accuracy once usable capacity is exhausted, which is consistent with the case-by-case failures reported in [3–5].

The converse has one exemption. A verifier that dereferences an external ground-truth store, checking candidates against the authoritative corpus rather than against the workspace, is not a read protocol over the workspace and is not covered by the bound. Whether a workspace-bound generator keeps *emitting* valid certificates as the workspace fills is a separate question; the experiment below tests it in comparative form.

The method. Definition 1 supplies the quantity, (3) makes it measurable, and inverting (3) gives the design parameters. Together they form the method:

1. *Calibrate the answer alphabet.* Construct a query battery in which every query admits exactly v equiprobable answers.

2. *Test grounding.* Run the identical prompt path over an *empty* workspace, including an arm that forces a guess, since refusals alone are no evidence of grounding. If accuracy exceeds the chance level $1/v$, the reader is answering from parametric memory, the Markov chain fails, and the method does not apply.

3. *Load progressively.* For each load ℓ on a chosen grid, render ℓ atomic facts into the workspace in a single fixed permutation, with target facts at fixed depth percentiles so that position is controlled, and run the battery under the read protocol whose capacity is being determined.

4. *Measure the error.* Compute the mean per-query error $\bar{\varepsilon}(\ell)$ under an all-trials convention fixed before measurement: a response that does not parse into a valid answer counts as an error, because information a workflow cannot emit is not usable information.

5. *Form the per-item usable information,*

$$\iota(l) \;=\; (h_0 - \varphi_v\left(\underline{\varepsilon}(l)\right))_+. \qquad (4)$$

6. *Bound the capacity from below.* Take $\hat{I}(\ell) = \ell\ \iota(\ell)$, which by (3) estimates a lower bound on $C_{\mathrm{ws}}(m;\ \mathscr{B}_{\mathrm{task}})$, and let $\hat{I}_{\max} = \max_{\ell} \hat{I}(\ell)$ be the point estimate. At each cell, take the one-sided 95% Clopper–Pearson upper confidence limit on $\bar{\varepsilon}(\ell)$ under worst-case full clustering, at one independent observation per query, and map it through (4); since (4) decreases in ε, the result is a one-sided 95% lower confidence bound at that load. $\hat{C}_{\mathrm{ws}}$ is the largest of these bounds, reported with the load attaining it.

7. *Invert for the design parameter.* For a target error $\varepsilon_0 \leq 1 - 1/v$ with $\varphi_v(\varepsilon_0) < h_0$, the greatest load at which the lower confidence bound meets the capacity requirement is

$$l_{max}^{(\varepsilon)} \;=\; \frac{\hat{C}_{ws}(m; B_{task})}{h_0 - \varphi_v(\varepsilon_0)}. \qquad (5)$$

8. *Invert for the coverage-limited load.* Where candidates carry a verifiable certificate, the same bound yields a second design parameter, given by (6) below with $\hat{C}_{\mathrm{ws}}$ in place of C_{ws}.

Two properties make the output usable in design. Because $\hat{C}_{\mathrm{ws}}$ is a lower confidence bound, both loads are lower confidence bounds as well: with the same confidence, the reported boundary lies at or below the true one, so the region it delimits is contained in the region the converse leaves open. And because the envelope is topology-independent at a fixed read budget and protocol class, the values apply unchanged to any orchestration over the same workspace that uses the same budget and class, so a change of aggregation scheme within that class does not require remeasurement.

Condition of applicability. Steps 4–6 attribute the whole of $\bar{\varepsilon}(\ell)$ to the read channel, which holds only where the error is retrieval-limited. Per-candidate accuracy factors exactly into a protocol-compliance term and a retrieval term, $\alpha = \kappa \cdot R$, where κ is the probability that a candidate emits a parseable certificate and R the accuracy given one. Where $\kappa = 1$ the per-item usable information ι measures the read channel; where κ is below 1, ι mixes the output grammar into the

estimate. The method therefore includes a diagnostic: compute the factorization at every load and interpret ι in its light. On the grid reported below, $\hat{C}_{ws}$ is attained at a cell with $\kappa = 1.000$, so it carries no compliance component.

Certificates convert error into abstention. Where candidates carry verifiable certificates, the natural policy answers when a certificate is valid and abstains otherwise. Write $\alpha(\ell)$ for per-candidate accuracy, $\gamma(\ell)$ for certificate completeness (the probability that a correct candidate is certified), $\delta(\ell)$ for the soundness leak (the probability that an incorrect one is), and $f(\ell)$ for coverage, the fraction of candidates released as answers. All four quantities are per-candidate.

Proposition 2 (Certified risk–coverage). *The selective error odds factor as*

$$\frac{1-\alpha}{\alpha} \cdot \frac{\delta}{\gamma}.$$

Under a sound certifier ($\delta = 0$) the selective risk is zero at every load and $f(\ell) = \alpha(\ell)\gamma(\ell)$, so the whole capacity penalty is transferred from error to coverage; sustaining $f(\ell) \geq f_0$, for a target with $\varphi_v(1 - f_0/\sup_\ell \gamma(\ell)) < h_0$ and $0 \leq 1 - f_0/\sup_\ell \gamma(\ell) \leq 1 - 1/v$, then requires

$$l \leq \frac{C_{ws}(m; B_{task})}{h_0 - \varphi_v(1 - f_0/\gamma(l))}. \qquad (6)$$

The gate divides the error odds by the certificate's positive likelihood ratio γ/δ, so a capacity limit need not be paid in wrong answers. The risk–coverage trade-off is itself standard in selective prediction [15]; the new element here is its coupling to a capacity converse, which bounds the achievable coverage of every orchestration over the shared workspace by the same C_{ws}. Because a true δ of zero can be bounded above but never observed, the measured form substitutes a one-sided upper confidence limit on δ. Since $f = \alpha\gamma$ and $\gamma \leq 1$, a coverage target is always at least as demanding as the numerically equal accuracy target.

Experimental setup. The corpus holds 4000 atomic facts in four-variant confusable clusters, so $v = 4$ and $h_0 = 2$ bits exactly. Fact values are randomized at corpus build, which makes B2 hold by construction. Queries are two-hop. Each candidate emits a certificate naming both facts: a READ line with the retrieved value and a LINK line citing the chain fact by identifier. The checker dereferences those identifiers against the authoritative corpus and never string-matches the model's prose, so it is an external-store verifier of the kind exempted above. Every run uses one reader model, `gemma-4-26b-a4b-it-maas`, at temperature 0.7, against one corpus pinned by SHA-256 and verified identical across every run. The protocol spends a read budget of three calls per query, one candidate per call, and gates the pool on the certificate. Hypotheses, decision rules, equivalence bands and estimands were fixed before the runs and not amended. All intervals are cluster-bootstrap intervals that resample queries: candidates within one pool share a byte-identical context and differ only in the generation draw, with measured within-pool correlation of 0.95–0.98, so they are never treated as three independent observations.

Results. In the grounding control over an empty workspace, the plain arm scores 0.0000 on 155 trials and the guess-forced arm 1/155 (0.0065), both far below the $v = 4$ chance level of 0.25, so the control gives no sign of a parametric side channel.

Steps 5–6 applied at each load give the per-item usable information and the capacity estimates, and the factorization gives the compliance diagnostic; all are collected in Table 1.

Table 1. Capacity estimates and the compliance diagnostic across the measured grid

ℓ	50	100	250	500	1000	2000	3000
$\bar{\varepsilon}(\ell)$	0.1475	0.0535	0.0258	0.0387	0.1634	0.3484	0.4581
$\kappa(\ell)$	0.8476	0.9373	0.9878	0.9971	1.0000	0.9978	0.9971
$R(\ell)$	0.9993	0.9990	0.9848	0.9583	0.8301	0.6487	0.5327
$\iota(\ell)$ (bits/item)	1.1628	1.6144	1.7862	1.7023	1.0985	0.5152	0.2791
$\hat{I} = \ell\iota$ (bits)	58.14	161.44	446.55	851.15	1098.47	1030.40	837.19
ceiling ℓh_0 (bits)	100	200	500	1000	2000	4000	6000

Here $\bar{\varepsilon}$ is the error of the workflow's final answer, from which ι follows by (4), while κ and R are per-candidate, so their product is per-candidate accuracy and not $1 - \bar{\varepsilon}$.

The four loads $\ell \le 500$ come from a downward extension of the grid, run separately with the same reader model, temperature, corpus (by SHA-256), 155-query battery, permutation seed, prompts and certificate grammar. The one difference is that the two lowest cells place their target facts at seven position percentiles rather than three; restricted to the three strata every cell shares, $\kappa(50) = 0.8681$ and $\kappa(100) = 0.9534$, still rising and still below every higher load.

Below $\ell = 250$ the compliance term dominates the error. In a follow-up run at $\ell = 50$ over the three shared strata (the main run did not keep raw model text), all 172 non-emitting candidates failed on one branch of the certificate grammar: the LINK line omitted the literal keyword cites. Supplying the keyword made all 172 parse and score correct. This count is separate from the cell rate in Table 1. The estimand was not revised: non-emitting candidates remain errors in every reported statistic. The decomposition shows that at those cells ι is not a retrieval measurement; the bound is taken at a load where compliance is complete.

The point estimate $\hat{I}_{\max}$ is 1098.47 bits, at $\ell = 1000$, interior to the measured grid (Fig. 1). The interiority is arithmetic: since $\hat{I} = \ell\iota \le \ell h_0$, every cell below 549 facts lies under 1098.47 bits by the ceiling alone, which covers all four low-load cells. The decline above $\ell = 1000$ is measured, but its interval includes zero: $\hat{I}(1000) - \hat{I}(3000) = 241.0$ bits, with a 95% interval of $[-50.8, +507.1]$. That interval comes from the one analysis that bootstraps this contrast, which scores each query by the correctness flag recorded during the run instead of re-deriving the plurality tie-break from the raw outputs; the two scorings differ by 0.00215 in $\bar{\varepsilon}$ at each end, in opposite directions, which accounts for the 20-bit gap to the difference computed from Table 1. The largest one-sided 95% lower confidence bound, attained at $\ell = 1000$ under worst-case full clustering, is $\hat{C}_{\mathrm{ws}} = 890.12$ bits, at a level that holds per load rather than simultaneously over the grid; the design parameters are computed from it.

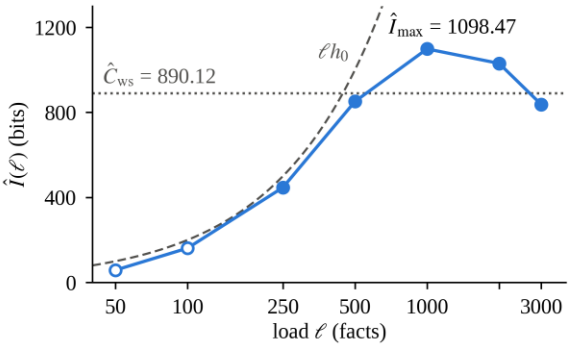


Fig. 1. Usable information $\hat{I}(\ell)$ across the measured grid, with the ceiling ℓh_0 and the lower confidence bound $\hat{C}_{\mathrm{ws}}$. Open markers: $\ell < 250$, where the compliance term dominates the error.

Applying steps 7–8 gives the method's output. At a target per-query accuracy of 0.75, (5) returns $l_{max}^{(\varepsilon)} = 1123$ facts; at a target coverage of 0.75, (6) returns $l_{max}^{(f)} = 1038$ facts, and the coverage constraint is necessarily the tighter of the two. Were (2) to hold with equality at the point estimate, $\hat{I}_{\max}$ would put the 0.75 accuracy crossing at 1098.47 / 0.7925 = 1386 facts, and the gap down to 1123 is the margin of the 95% lower confidence bound under worst-case clustering. The system's own accuracy meets the target at $\ell = 1000$ ($1 - \bar{\varepsilon} = 0.8366$) and misses it at $\ell = 2000$ (0.6516), so its crossing lies between these two measured loads, as do both 1386 and 1123. The grid does not locate the crossing more finely, and the comparison does not test the converse. Against the nominal ceiling the gap is large: at $\ell = 1000$ the workspace carries 555,442 bits with a vocabulary of about 256k tokens, so $\hat{C}_{\mathrm{ws}}$ is 1/624 of it.

Across $\ell = 1000, 2000, 3000$, accuracy α falls $0.8301 \to 0.6473 \to 0.5312$ while certifier completeness γ falls $0.9741 \to 0.9214 \to 0.9177$. The primary contrast is between the two declines: $\Delta\alpha - \Delta\gamma = 0.2425$ with a one-sided 95% lower bound of 0.1883, above the pre-stated floor of 0.15. Generation degrades materially more than certification does.

The second test asks whether γ's decline is *equivalent to zero* within a band of ± 0.075

fixed before the runs. It is inconclusive: the trend estimate is 0.0652 with a 90% interval of [0.0334, 0.0962], straddling a band boundary. That interval excludes zero, so γ declined significantly, and the decoupling is established in its comparative form and not as an invariance of the checker.

Across 1383 incorrect candidates pooled over the three loads the certifier admitted none, giving $\delta = 0$ with a Wilson upper bound of 0.0054 on the measured design-effect basis and 0.0082 under worst-case fully-clustered correction, against a pre-stated 0.05. Precision on certified answers is 1.000 at every load while coverage falls 0.8086 → 0.5964 → 0.4875: under a sound certifier the capacity limit shows up as a loss of coverage and not of correctness.

A degraded-certifier ladder run at $\ell = 2000$ on 492 of those incorrect candidates shows that the checker can register leaks, so the zero is not an artifact of an insensitive instrument: weakening the checker produces leaks of 0.0305, 0.0224, 0.9146 and 0.9939 across four rungs, every interval excluding zero.

Scope. This work uses one reader model, one synthetic corpus and one certificate grammar. Within that scope capacity is bounded *below* only: the estimator gives lower confidence bounds, and nothing here produces an upper bound at any load. The method applies to batteries calibrated so that every query admits exactly v equiprobable answers and to readers that pass the empty-workspace grounding test, which excludes deployments in which the model already knows the facts. The measure is indexed to the read budget and protocol class, so a value measured at one budget is not guaranteed to hold at another. Topology invariance is derived from the converse; no experiment here varies the topology. The vocabulary size $|\Sigma| \approx 256$k behind the 1/624 ratio is an assumption; halving or doubling it moves the nominal figure by about one bit per token.

Conclusions

We have developed a method for determining the maximum admissible semantic load on a shared LLM context for a required answer accuracy, based on a usable capacity defined at the read interface and a Fano-type converse for it. From black-box API calls alone, the method determines the admissible load and states the minimum capacity requirement a model must meet for a given load and accuracy; below the admissible load, an accuracy shortfall lies in the read strategy rather than in the context. For the reader model studied, the lower confidence bound on capacity was nearly three orders of magnitude below nominal capacity, and the admissible load for the accuracy target lies between the largest measured load at which the system met that target and the smallest at which it did not.

With verifiable certificates, the method gives a second admissible load, for a target coverage. Under a sound certifier the capacity limit shows up as a loss of coverage and not of correctness: every certified answer was correct at every load, and the accuracy of generated candidates declined by more than the certifier's completeness did.

At a fixed read budget and protocol class, the measure does not depend on the topology of the multi-agent system: the converse gives one envelope for every orchestration of agents reading the shared context, from a single agent to voting and verification schemes.

It follows from the converse that at a fixed read budget and protocol class, for orchestrations that locate answers in the context no aggregation scheme or topology redesign restores accuracy once usable capacity is exhausted. The proposed method makes it possible to set minimum requirements on the model when such systems are designed.

Use of AI Tools

During the preparation of this work the authors used AI-based tools for formatting and stylistic revision of the manuscript and for technical assistance in writing the data-collection and analysis code. The reader model under study, gemma-4-26b-a4b-it-maas, is the object of measurement and not an authoring tool. No AI-generated text is presented as a scientific result, no data, results or sources were

generated by such tools, and no tool is named as an author. All scientific ideas, the design of the experiments, the interpretation of the measurements and the conclusions are the authors' own.

References

1. Wang X., Wei J., Schuurmans D. et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *Proceedings of the International Conference on Learning Representations (ICLR 2023)*. 2023. DOI: 10.48550/arXiv.2203.11171

2. Li J., Zhang Q., Yu Y., Fu Q., Ye D. More Agents Is All You Need. *Transactions on Machine Learning Research*. 2024. DOI: 10.48550/arXiv.2402.05120

3. Chen L., Davis J. Q., Hanin B. et al. Are More LLM Calls All You Need? Towards the Scaling Properties of Compound AI Systems. *Advances in Neural Information Processing Systems (NeurIPS 2024)*. 2024. DOI: 10.52202/079017-1455

4. Kim Y., Gu K., Park C. et al. Towards a Science of Scaling Agent Systems. *arXiv preprint arXiv:2512.08296*. 2025. DOI: 10.48550/arXiv.2512.08296

5. Cemri M., Pan M. Z., Yang S. et al. Why Do Multi-Agent LLM Systems Fail? *arXiv preprint arXiv:2503.13657*. 2025. DOI: 10.48550/arXiv.2503.13657

6. Yang Y., Qu C., Wen M. et al. Understanding Agent Scaling in LLM-Based Multi-Agent Systems via Diversity. *arXiv preprint arXiv:2602.03794*. 2026. DOI: 10.48550/arXiv.2602.03794

7. Du Y., Tian M., Ronanki S. et al. Context Length Alone Hurts LLM Performance Despite Perfect Retrieval. *Findings of the Association for Computational Linguistics: EMNLP 2025*. 2025. DOI: 10.18653/v1/2025.findings-emnlp.1264

8. Liu N. F., Lin K., Hewitt J. et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. 2024. Vol. 12. P. 157–173. DOI: 10.1162/tacl_a_00638

9. Paulsen N. Context Is What You Need: The Maximum Effective Context Window for Real World Limits of LLMs. *Advances in Artificial Intelligence and Machine Learning*. 2026. Vol. 6, no. 1. P. 4853–4878. DOI: 10.54364/AAIML.2026.61268

10. Wan K., Gao L., Mu H. et al. A Fano-Style Accuracy Upper Bound for LLM Single-Pass Reasoning in Multi-Hop QA. *Proceedings of the International Conference on Learning Representations (ICLR 2026)*. 2026. DOI: 10.48550/arXiv.2509.21199

11. Xu Y., Zhao S., Song J., Stewart R., Ermon S. A Theory of Usable Information Under Computational Constraints. *Proceedings of the International Conference on Learning Representations (ICLR 2020)*. 2020. DOI: 10.48550/arXiv.2002.10689

12. Stroebl B., Kapoor S., Narayanan A. The Limits of Inference Scaling Through Resampling. *arXiv preprint arXiv:2411.17501*. 2024. DOI: 10.48550/arXiv.2411.17501

13. Stechly K., Valmeekam K., Kambhampati S. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. *Proceedings of the International Conference on Learning Representations (ICLR 2025)*. 2025. DOI: 10.48550/arXiv.2402.08115

14. Farquhar S., Kossen J., Kuhn L., Gal Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*. 2024. Vol. 630. P. 625–630. DOI: 10.1038/s41586-024-07421-0

15. Geifman Y., El-Yaniv R. Selective Classification for Deep Neural Networks. *Advances in Neural Information Processing Systems (NIPS 2017)*. 2017. DOI: 10.48550/arXiv.1705.08500

Рукопис опрацьовував незалежний відповідальний редактор; автор не брав участі у процесі рецензування та прийнятті рішення щодо публікації.

Tachinina O., Arutiunian A.

METHOD FOR DETERMINING THE MAXIMUM SEMANTIC LOAD ON A SHARED CONTEXT FOR A REQUIRED ANSWER ACCURACY

We propose a method for determining the maximum semantic load that a context shared by LLM-based agents can carry while still admitting a required probability of a correct answer. The method rests on usable capacity, defined at the read interface for a given read budget and class of read protocols, and on a Fano-type converse whose bound does not depend on the orchestration topology. Applied to the error measured through the model's API alone, without access to its internal state, the converse yields a lower confidence bound on usable capacity. This bound sets a minimum capacity requirement on the model and gives a conservative admissible load, in atomic facts, for a target accuracy. When candidate answers carry verifiable certificates, the same bound gives an admissible load for a target fraction of queries answered. For one reader model on synthetic two-hop queries, the lower confidence bound on capacity was nearly three orders of magnitude below nominal capacity, and the admissible load for the accuracy target lay between the largest measured load that met the target and the smallest that missed it. When answers were released only after their certificate passed verification, every certified answer was correct, and the capacity limit appeared as a falling fraction of queries answered. By the converse, in systems that locate answers in the context, at a fixed read budget and protocol class, no aggregation scheme or change of topology restores accuracy once usable capacity is exhausted, so the method lets requirements on the model be set at the design stage.

Keywords: *certificate verification; channel capacity; context window; Fano inequality; multi-agent systems; selective prediction*

Тачиніна О., Арутюнян А.

МЕТОД ВИЗНАЧЕННЯ МАКСИМАЛЬНОГО СЕМАНТИЧНОГО НАВАНТАЖЕННЯ СПІЛЬНОГО КОНТЕКСТУ ЗА ЗАДАНОЇ ТОЧНОСТІ ВІДПОВІДІ

У статті запропоновано метод визначення максимального семантичного навантаження спільного контексту агентів на основі великої мовної моделі, яке допускає досягнення заданої ймовірності правильної відповіді. Метод ґрунтується на корисній пропускній здатності, означеній на інтерфейсі читання за заданих бюджету і класу протоколів читання, та на оберненій теоремі типу Фано, межа якої не залежить від топології оркестрації. Застосування теореми до похибки, виміряної лише через API моделі без доступу до її внутрішнього стану, дає нижню довірчу межу корисної пропускної здатності. Ця межа задає мінімальну вимогу до пропускної здатності моделі та консервативне допустиме навантаження для цільової точності, виражене кількістю атомарних фактів. Якщо відповіді супроводжуються сертифікатами, які можна верифікувати, та сама межа дає допустиме навантаження для цільової частки наданих відповідей. Для однієї моделі-читача на синтетичних дволанкових запитах нижня довірча межа корисної пропускної здатності виявилася майже на три порядки меншою за номінальну пропускну здатність, а допустиме навантаження для цільової точності лежить між найбільшим виміряним навантаженням, за якого цю точність досягнуто, і найменшим, за якого її не досягнуто. Коли відповідь видавалася лише після успішної перевірки сертифіката, усі сертифіковані відповіді були правильними, а обмеження пропускної здатності проявилося в зменшенні частки наданих відповідей. З оберненої теореми випливає, що в системах, які знаходять відповіді в контексті, за фіксованих бюджету і класу протоколів читання жодна схема агрегації чи зміна топології не здатна відновити точність після вичерпання корисної пропускної здатності, тож метод дає змогу задати вимоги до моделі ще на етапі проєктування.

Ключові слова: *багатоагентні системи; верифікація сертифікатів; контекстне вікно; нерівність Фано; пропускна здатність каналу; селективне прогнозування*

Стаття подана до редакції: 11*/09/2026*
Стаття прийнята до опублікування:18/09/2026
Стаття опублікована: 10/11/2026

Стаття поширюється на умовах ліцензії CC BY 4.0