ENTROPIC CONTEXT SHAPING: INFORMATION-THEORETIC FILTERING FOR CONTEXT-AWARE LLM AGENTS

Hyunjun Kim
KAIST
hyunjun1121@kaist.ac.kr

ABSTRACT

Context engineering for large language model (LLM) agents requires distinguishing pragmatically useful information from misleading distractors. We introduce En- tropic Context Shaping (ECS), an information-theoretic framework that measures context utility via the shift in the model's answer distribution toward the correct answer. Unlike lexical similarity methods that rely on word overlap, ECS captures pragmatic utility–whether a passage actually helps answer the question. We formalize utility as the signed change in answer probability and provide theoretical analysis showing that task-irrelevant updates yield near-zero distribution shift. We evaluate on multi-turn context selection tasks using LongMemEval (session-level) and LoCoMo (turn-level) benchmarks. On fine-grained turn selection, ECS with Llama-3.1-8B achieves F1=0.265, a 72% relative improvement over TF-IDF (F1=0.154), demonstrating that pragmatic utility outperforms lexical similarity when precise context selection matters.

1 INTRODUCTION

Large language model (LLM) agents increasingly rely on dynamically constructed context windows to maintain state and inform decision-making (Yao et al., 2023; Shinn et al., 2023). As these agents interact with complex environments, they must continuously decide which observations, memories, and intermediate results to retain. This process–context engineering–has emerged as a critical bottleneck in agent performance. The Accumulation Fallacy. Current approaches to context management typically employ semantic similarity metrics (e.g., cosine distance between embeddings) to filter redundant information (Lewis et al., 2020; Guu et al., 2020). While effective at removing near-duplicates, these methods suffer from a fundamental flaw: they conflate semantic novelty with pragmatic utility. A piece of information may be entirely unique yet completely irrelevant to the task at hand. We term such distractors "Red Herrings"–facts that are true, novel, and semantically distinct, but provide zero information gain for the downstream decision. This distinction between semantic and pragmatic utility is critical for long-horizon agent tasks. As agents accumulate observations over extended interactions, purely semantic filtering allows context to grow with irrelevant but unique facts. The resulting context dilution degrades performance: models struggle to attend to genuinely useful information buried among distractors (Liu et al., 2024). Moreover, the computational cost of processing bloated contexts becomes prohibitive, as attention complexity scales quadratically with context length. Motivating Example. Consider an LLM agent solving a mathematical word problem. The follow- ing context updates arrive sequentially: 1. Insight: "The train's speed is 60 km/h." (Critical for solution) 2. Redundant: "The vehicle moves at sixty kilometers per hour." (Duplicate) 3. Red Herring: "The conductor has worked for 15 years and enjoys jazz." (Novel but useless) Semantic filtering correctly rejects (2) as redundant but accepts (3) because it is semantically unique. This is the Accumulation Fallacy in action. Our Contribution: Entropic Context Shaping (ECS). We propose an information-theoretic framework that measures the pragmatic utility of context updates by their impact on the model's answer distribution. Specifically, we:

• Define utility as the shift in the model's output distribution toward the correct answer–capturing whether a passage helps vs. misleads (Section 3) • Provide theoretical analysis showing that task-irrelevant updates yield near-zero distribution shift under mild assumptions (Theorem 3.2) • Demon-

strate that ECS achieves 72% relative improvement over TF-IDF on fine-grained turn-level selection (LoCoMo), while identifying boundary conditions where semantic methods remain superior (LongMemEval session-level) • Analyze model dependence, showing 13× performance variation between Llama-3.1-8B and Qwen2.5-7B on identical data–highlighting the importance of model-specific validation for logprob-based approaches Key Findings. Our experiments on multi-turn context selection reveal three insights: (1) ECS excels at fine-grained selection where semantic overlap is uninformative, achieving F1=0.265 versus TF-IDF's 0.154 on turn-level evidence selection; (2) ECS performance is strongly model-dependent, suggesting that logprob informativeness varies substantially across architectures; and (3) at coarse granularity (session-level selection), semantic methods dominate, establishing an important boundary condition for ECS applicability. These findings suggest a hybrid approach where semantic filtering handles coarse candidate reduction while ECS performs fine-grained ranking.

2 Related Work

Context Engineering for LLM Agents. The challenge of managing context in LLM agents has received significant attention (Yao et al., 2023; Shinn et al., 2023; Wei et al., 2022). Most approaches treat context as a retrieval problem, selecting relevant documents or memories based on semantic similarity (Lewis et al., 2020; Karpukhin et al., 2020). While effective for information retrieval, these methods are ill-suited for agent contexts where the goal is task completion rather than document relevance. Recent work has shown that LLMs exhibit position-dependent biases when processing long contexts, often failing to utilize information in the middle of the input (Liu et al., 2024). The evolution from "prompt engineering" to "context engineering" reflects a growing recognition that effective LLM deployment requires not just careful prompt design, but systematic management of the information that informs model predictions. This shift is particularly critical for agentic applications where contexts accumulate over extended interactions and must be curated to maintain coherent task-relevant knowledge. Retrieval-Augmented Generation (RAG). RAG systems (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022) improve LLM performance by conditioning on retrieved documents. However, RAG assumes a static document corpus, whereas agent contexts evolve dynamically. Furthermore, RAG's reliance on embedding similarity inherits the Accumulation Fallacy we identify. Self-RAG (Asai et al., 2024) addresses when to retrieve through self-reflection tokens, but does not address what retrieved content is pragmatically useful. Recent work on RAG robustness (Fang et al., 2024; Chen et al., 2024) addresses noise in retrieved documents through adversarial training and benchmarking, but does not resolve the fundamental semantic-pragmatic mismatch. Context compression approaches (Jiang et al., 2024; Cheng et al., 2024) reduce token count while preserving information, but do not filter based on task utility. Recent work on RAG evaluation (Es et al., 2024) provides reference-free metrics for assessing retrieval quality, complementing our approach by offering tools to measure pragmatic utility in practice. Memory Systems for Agents. Recent work on agent memory (Park et al., 2023; Wang et al., 2023) has explored hierarchical and episodic memory structures. A comprehensive survey (Zhang et al., 2024) identifies memory as the key component transforming LLMs into "true agents." These systems typically use recency and importance heuristics for memory consolidation. Our work complements these approaches by providing a principled utility metric for individual memory updates. Red Herrings and Inconsequential Noise. The phenomenon of misleading information in rea- soning tasks has been studied extensively. The Only Connect Wall (OCW) dataset (Taati et al., 2023) demonstrated that LLMs are "fixated" by Red Herrings, failing to ignore irrelevant but salient information. Recent work on counterfactual robustness (Liu et al., 2023) shows that LLMs are susceptible to interference from unreliable external knowledge. We build on these observations by proposing an information-theoretic solution that distinguishes pragmatically useful context from semantically novel but useless distractors.

Information-Theoretic Perspectives on LLMs. Prior work has applied information theory to analyze LLM behavior (Xu et al., 2020; Ethayarajh et al., 2022). Voita et al. (2019) demonstrated that information flow through transformer layers can be analyzed via probing tasks, revealing which layers encode different linguistic properties. Work on minimum description length (Voita & Titov, 2020) has shown that attention patterns correlate with information compression. Our approach differs by using KL-divergence as an operational filter rather than a diagnostic tool, and by extending to multi-step trajectory divergence for improved separation. The key conceptual shift is from using information- theoretic measures to understand LLM behavior to using them to control what informa-

tion influences predictions. Long-Context LLMs and Attention Mechanisms. Recent advances have extended LLM context windows from 2K to over 128K tokens (Chen et al., 2023; Peng et al., 2023), yet longer contexts do not automatically improve task performance. Studies show that model performance degrades with context length even when relevant information is present (Liu et al., 2024), suggesting that attention mechanisms struggle to identify task-relevant tokens amid large amounts of irrelevant content. Efficient attention mechanisms like Flash Attention (Dao et al., 2022) and sliding window approaches (Beltagy et al., 2020) address computational constraints, while recent work on attention optimization (Hooper et al., 2024; Pan et al., 2025) accelerates long-context inference through KV-cache compression. However, these approaches do not resolve the fundamental question of what should be in the context. Our work complements these architectural advances by providing a principled method to curate context content, ensuring that extended context windows contain information that genuinely aids task completion.

3 METHODOLOGY

We formalize context filtering as an information-theoretic decision problem. Given an evolving context C and a candidate update u, we seek to determine whether u provides sufficient pragmatic utility to warrant inclusion.

3.1 PROBLEM FORMULATION

Let M be a language model with vocabulary V and let q be a query or task specification. For any context C, the model induces a distribution over next-token predictions: PC (y) = P (y — C, q; M) (1) The Direction Problem. A key insight is that measuring distribution change (magnitude) is insufficient–we must also assess whether the change moves the distribution toward the correct answer. Consider two passages for the query "Super Bowl 2021 location": • Correct: "...held in Tampa, Florida..." • Counterfactual: "...held in Glendale, Arizona..." (false) Both induce large distribution shifts (high KL-divergence), but only the first shifts toward the correct answer. Pure magnitude-based filtering would accept both.

3.2 PRAGMATIC UTILITY: THEORETICAL FRAMEWORK

We define pragmatic utility as the signed change in the model's belief about the correct answer, not merely the unsigned distribution shift: Definition 3.1 (Pragmatic Utility). The pragmatic utility of a candidate update u with respect to context C and ground truth answer a is: U(u; C, a) = log PCu (a) log PC (a) · —u— (2) where —u— denotes the token length of u, and 0 is a length penalty. This definition captures the intuitive notion that a passage is useful if it increases the model's probability of generating the correct answer. When ground truth is unavailable, we can approximate this using the LLM's parametric knowledge as a proxy (Section 3.5). Relationship to KL-Divergence. When the model's prior belief is concentrated on incorrect answers, useful passages induce large KL-divergence toward the correct answer. The magnitude of DKL (PCu PC) is a necessary but not sufficient condition for utility–we additionally require the direction to be correct.

3.3 RED HERRING THEOREM

The theoretical framework provides guarantees for task-irrelevant updates: Theorem 3.2 (Red Herring Rejection). Let u be task-irrelevant (i.e., u q given C). Under regularity conditions on M: DKL (PCu PC) (3) for small ¿ 0 depending on model capacity. Proof Sketch. Task-irrelevance implies conditional independence between u and the answer given C. By the data processing inequality (Cover & Thomas, 2006), adding u cannot increase mutual information with the answer. Scope and Limitations. Theorem 3.2 addresses "red herrings"–semantically unique but pragmati- cally useless passages. It does not address counterfactual passages that contain plausible but incorrect information. Detecting such passages requires direction-aware scoring, which we operationalize in Section 3.5.

3.4 MULTI-TOKEN TRAJECTORY DIVERGENCE

Single-token KL-divergence can fail when the first token is generic (e.g., ”The”, ”Step”). We extend to multi-token trajectory divergence: Definition 3.3 (Trajectory Divergence). For a generation horizon T , the trajectory divergence is: T (t) (t) X DT (u; C) = DKL (PCu PC) (4) t=1 where P (t) denotes the distribution at generation step t. This cumulative measure amplifies the signal from useful passages while irrelevant passages maintain near-zero divergence across all steps.

3.5 OPERATIONALIZING PRAGMATIC UTILITY

Definition 3.1 requires access to PC (a)–the model's probability assigned to the correct answer. We operationalize this in two settings: Setting 1: Ground Truth Available. When the correct answer a is known (e.g., during training data curation, retrieval system evaluation, or active learning), we can directly compute Equation 2 via logprob extraction, or use an LLM-as-Judge (Zheng et al., 2023) that evaluates: ”Does this passage help answer the question correctly?” Setting 2: Ground Truth Unavailable. At inference time, when a is unknown, we leverage the LLM's parametric knowledge. The factuality-aware scoring asks: ”Based on your knowledge, does this passage contain accurate information for answering this question?” This approximates the direction-aware criterion by using the model's prior knowledge to verify passage correctness. Algorithm. Algorithm 1 summarizes our approach. In the ground-truth-available setting, the utility function directly measures answer probability change. In the ground-truth-unavailable setting, we use factuality-aware LLM scoring as a proxy.

3.6 WORKED EXAMPLE: ECS SCORING IN PRACTICE

To illustrate how ECS distinguishes useful context from red herrings, consider a concrete example with the question: ”What year was the Eiffel Tower completed?” Setup. Let C0 = (empty initial context). Without any context, the model's distribution over the first answer token might be: PC0 (”1889”) = 0.35 PC0 (”1887”) = 0.20 PC0 (”1890”) = 0.15 PC0 (other) = 0.30 The model has some parametric knowledge but assigns probability mass across several candidate years.

Algorithm 1 Entropic Context Shaping (ECS) Require: Context C, candidate u, query q, threshold Require: Optional: ground truth a (if available) 1: if a is available then 2: U ← log PCu (a) log PC (a) Answer-aware scoring 3: else 4: U ← FactualityScore(q, u; M) Factuality-aware proxy 5: end if 6: if U ¿ then 7: Accept: C ← C u 8: else 9: Reject: discard u 10: end if 11: return C Candidate 1: Useful Context. Consider u1 : ”The Eiffel Tower was inaugurated on March 31, 1889, after two years of construction.” Adding this context shifts the distribution sharply: PC0 u1 (”1889”) = 0.92 PC0 u1 (”1887”) = 0.03 PC0 u1 (”1890”) = 0.02 PC0 u1 (other) = 0.03 The pragmatic utility (assuming correct answer is 1889) is: U(u1) = log(0.92) log(0.35) = 0.97 nats Candidate 2: Red Herring. Now consider u2 : ”Gustave Eiffel also designed the internal framework for the Statue of Liberty.” This fact is semantically related (mentions Eiffel) but does not help answer the question about completion year: PC0 u2 (”1889”) = 0.38 PC0 u2 (”1887”) = 0.19 PC0 u2 (”1890”) = 0.14 PC0 u2 (other) = 0.29 The pragmatic utility is minimal: U(u2) = log(0.38) log(0.35) = 0.08 nats ECS Decision. With a threshold = 0.3 nats, ECS accepts u1 (utility 0.97 ¿) and rejects u2 (utility 0.08 ¡). Importantly, a semantic similarity method like cosine similarity might rank u2 highly due to its lexical overlap with ”Eiffel,” demonstrating why pragmatic scoring outperforms semantic matching for fine-grained selection. Multi-Token Extension. The advantage becomes more pronounced with trajectory divergence. If the full answer is ”1889, after roughly two years of construction,” the useful context u1 maintains high alignment across multiple tokens, while u2 's influence remains negligible–amplifying the separation between useful and irrelevant context.

3.7 COMPUTATIONAL EFFICIENCY VIA PREFIX CACHING

When using logprob-based scoring, we leverage prefix caching (Kwon et al., 2023; Xiao et al., 2024) to achieve near-constant overhead: Proposition 3.4 (Amortized Complexity). With prefix caching, the amortized cost per candidate is O(—u—) rather than O(—C— + —u—), achieving O(1) with respect to context length. This is achieved by caching the key-value pairs for C, such that only the incremental tokens from u require computation.

3.8 When Does ECS Apply?

ECS is most effective in the following scenarios: • Retrieval evaluation: Comparing retrieval systems' ability to surface useful vs. misleading passages (ground truth available from benchmark labels). • Training data curation: Filtering context for instruction tuning or RLHF where correct answers are known. • Active learning: Prioritizing which retrieved passages to verify when labeling resources are limited.

Table 1: Context selection results across datasets and models (F1 scores ± std). LongMemEval: session-level selection (100 samples). LoCoMo: turn-level evidence selection (200 QA pairs). Bold indicates best method per column. High standard deviations reflect per-sample variance typical of retrieval tasks. LongMemEval (Sessions) LoCoMo (Turns) Method Qwen2.5-7B Llama-3.1-8B Qwen2.5-7B Llama-3.1-8B ECS (Ours) .152±.31 .139±.31 .020±.12 .265±.41 TF-IDF .649±.43 .649±.43 .154±.32 .154±.32 Dense (SBERT) .591±.46 .591±.46 –† –† Random .024±.12 .024±.12 .003±.02 .003±.02 † Dense retrieval omitted for LoCoMo due to computational cost at turn-level granularity. • Counterfactual robustness testing: Evaluating LLM robustness to factually incorrect context. For real-time inference without ground truth, the factuality-aware variant provides a practical approx- imation, though with reduced accuracy (Section 4).

4 Experiments

We evaluate ECS on multi-turn context selection tasks, measuring its ability to identify pragmatically useful information in long conversational histories.

4.1 Experimental Setup

Datasets We evaluate on two datasets with different selection granularities: • LongMemEval Wu et al. (2025): Session-level selection. Given a question and $\sim$ $50 conversation sessions (each $ \sim $6K tokens), select session (s) containing the answer. We use 100 samples with ground truth answer session ids. LoCoMo Turn − level evidence selection. Given a question about a long conversation ($ \sim $400 turns across multiple sessions), identify specific dialogue turns serving as evidence. We evaluate on 200 QA pairs fr LongMemEval tests coarse−grained session selection where semantic overlap is informative, while LoCoMo tests fine grained turn selection requiring pragmatic reasoning about which specific utterances contain evidence. Models We evalu tuned LLMs of similar scale : Qwen2.5 − 7B − Instruct Yang et al.(2024) : 7B parameter model with 8K context Llama − 3.1 − 8B − Instruct Grattafiori et al.(2024) : 8B parameter model with 8K context Using two models enables us to study model − dependence of the log prob−based approach−−a critical consideration for practical deployment. Baselines We compare E TF − IDF (Spärck Jones, 1972) : Cosine similarity between question and context TF − IDF vectors − −a strong lexical baseline Dense : Sentence − BERT (Reimers & Gurevych, 2019) (all − MiniLM − L6 − v2) embedding cosine similarity − −representing neural semantic matching Random : Uniform random selection baseline Metric We report F1 score for selecting k items matching the ground truth size. Since

4.2 Results

Table 1 presents the main results. Figure 1 visualizes the performance comparison across both datasets with error bars indicating standard deviation. Main Findings Our experiments reveal three key findings: 1. ECS excels at fine-grained selection. On LoCoMo (turn-level), ECS with Llama-3.1-8B achieves F1=0.265, a 72% relative improvement over TF-IDF (F1=0.154). This validates ECS's core

1.0 LongMemEval (Session Selection) 0.5 LoCoMo (Turn Evidence Selection) 0.8 0.4 0.65 +72% 0.59 0.6 0.3 0.27 F1 Score

F1 Score 0.4 0.2 0.15 0.2 0.15 0.14 0.1 0.02 0.02 0.00

4.3 ECS ECS TF-IDF DENSE RANDOM 0.0 ECS ECS TF-IDF RANDOM

(Qwen) (Llama) (Qwen) (Llama) Method Method Figure 1: Performance comparison across datasets and methods. Left: LongMemEval (session-level selection)–semantic methods (TF-IDF, Dense) dominate. Right: LoCoMo (turn-level selection)– ECS with Llama-3.1-8B achieves 72% relative improvement over TF-IDF. Error bars show standard deviation across samples. The contrasting results highlight the importance of matching method to granularity. hypothesis: measuring answer distribution shift via KL divergence captures pragmatic utility that semantic similarity misses. The improvement is particularly notable given that TF-IDF is a strong baseline for evidence retrieval. 2. ECS is model-dependent. Qwen2.5-7B shows poor ECS performance (F1=0.020 on LoCoMo), while Llama-3.1-8B succeeds (F1=0.265). This 13× difference on identical data suggests that model logprob distributions vary substantially in their informativeness for context utility scoring. We analyze this phenomenon in Section 4.3.2. 3. Coarse selection favors semantic methods. On LongMemEval (session-level), TF-IDF (F1=0.649) and Dense retrieval (F1=0.591) significantly outperform ECS (F1\$$\sim$ \$0.15). $At session granularity(\$ \sim \$6K tokens), irrelevant content overwhelms the logprob signal. This establishes an important boundary condition for EC$

4.4 ANALYSIS

We investigate two key phenomena: the granularity effect and model dependence.

4.4.1 GRANULARITY ANALYSIS

Figure 2 illustrates how ECS performance varies with context granularity. The pattern is striking: ECS improves as granularity becomes finer, while TF-IDF shows the opposite trend. Why does granularity matter? ECS computes KL(Pwith-context Pbase) for each candidate context. When contexts are entire sessions ($\$\sim$ $\$6K tokens), most content is irrelevant to the question, creating a low signal $-$ to $-$ noise ratio. The informative tokens that would shift the answer distribution are diluted by thousands of irrelevant tokens grained turn selection ($\$ \sim \$50-200 tokens per turn) allows ECS to isolate passages that directly shift answer probability use semantic methods for coarse filtering (reducing thousands of candidates to hundreds), then apply ECS for fine$-$ grained ranking. This two$-$stage pipeline leverages each method' s strengths while managing computational costs.$

4.4.2 MODEL DEPENDENCE ANALYSIS

Figure 3 shows the stark contrast between models on the same task. Why does model matter? The difference between Qwen (F1=0.020) and Llama (F1=0.265) suggests that model architecture significantly affects logprob informativeness. We hypothesize several contributing factors: • Tokenizer granularity: Different tokenizers produce different token boundaries, affecting which semantic units receive probability mass.

ECS Performance by Granularity (Llama-3.1-8B) 0.65 ECS

4.5 TF-IDF

0.5 0.4 F1 Score

0.3 0.27 0.2 0.15 0.14 0.1

4.6 SESSION TURN

(6K tokens) (50 tokens) Selection Granularity Figure 2: ECS performance by selection granularity (Llama-3.1-8B). At coarse session-level ($\$\sim$ $\$6K tokens), TF $-$ IDF dominates. At fine turn $-$ level ($\$ \sim \$50 tokens), ECS overtakes TF $-$ IDF. The dashed arrow shows ECS improving as granularity increases, suggesting pragmatic signals become detectable Model Dependence $ECS(Qwen2.5-7B)$

4.7 ECS (LLAMA-3.1-8B)

4.8 TF-IDF (BASELINE)

0.6 0.5 F1 Score

0.4 0.3 0.27 0.2 0.15 0.15 0.14 0.1 0.02

4.9 LONGMEMEVAL LOCOMO

(Sessions) (Turns) Dataset Figure 3: Model dependence of ECS performance. On LoCoMo, Llama-3.1-8B achieves F1=0.265 while Qwen2.5-7B achieves only F1=0.020–a 13× difference on identical data. TF-IDF (model- independent) provides consistent F1=0.154 baseline. This suggests logprob informativeness varies significantly across model architectures. • Pre-training distribution: Models trained on different corpora may develop different "sharpness" in their probability distributions. • Instruction tuning: The fine-tuning process may affect how confidently models assign probability to specific answers. Practical implications. This model-dependence has important implications: ECS effectiveness should be validated for each target model before deployment. However, the strong results with Llama suggest that ECS can achieve substantial improvements when paired with compatible models.

4.9.1 QUALITATIVE ANALYSIS

To understand when ECS succeeds and fails, we examine representative examples from LoCoMo. Success case. For the question "What restaurant did Alex recommend for the anniversary dinner?", ground truth evidence is turn 47: "Alex: I'd suggest trying Chez Laurent–their French cuisine is perfect for special occasions." ECS with Llama correctly ranks this turn highest because including it dramatically increases P ("Chez Laurent"—context). TF-IDF misranks turns containing generic food-related words ("dinner", "restaurant") that match the query but don't answer it. Failure case. ECS struggles when the answer requires reasoning across multiple turns. For questions like "How long has Sarah been learning piano?", the evidence spans turns 12 ("I started lessons in March") and turn 89 ("Can't believe it's been almost a year"). ECS evaluates turns independently and may miss such distributed evidence.

4.10 DISCUSSION

When to Use ECS vs. Semantic Methods. The contrasting results on LongMemEval and LoCoMo suggest a clear division of labor: use semantic methods (TF-IDF, dense retrieval) for coarse filtering among large chunks (¿1K tokens), and ECS for fine-grained ranking among focused passages (¡200 tokens). This motivates a two-stage pipeline combining both approaches. Model Architecture Effects. The difference between Llama (F1=0.265) and Qwen (F1=0.020) suggests that logprob informativeness depends on calibration, tokenizer alignment with answer boundaries, and training objectives. Understanding these factors would help predict model suitability for ECS without extensive testing. Implications for Long-Context LLMs. Even with 128K+ token windows, what goes into context matters as much as capacity (Liu et al., 2024). ECS provides principled curation ensuring extended windows contain genuinely useful information rather than noise. RAG System Design. Our granularity analysis has implications for retrieval-augmented generation: retrieving entire documents may introduce excessive noise, while fine-grained passages allow models to attend to relevant content. ECS can evaluate whether retrieved content improves answer quality beyond query matching.

5 CONCLUSION

We introduced Entropic Context Shaping (ECS), an information-theoretic framework for filtering context in LLM agents by measuring whether passages shift the model's answer distribution toward correct answers. Key Findings. (1) ECS excels at fine-grained selection: On LoCoMo turn-level selection, ECS with Llama-3.1-8B achieves F1=0.265, a 72% improvement over TF-IDF (F1=0.154). (2) Granular- ity matters: At coarse session-level (LongMemEval), semantic methods dominate; ECS requires fine-grained contexts. (3) Model dependence: Performance varies across models

(Llama F1=0.265 vs. Qwen F1=0.020), suggesting logprob informativeness depends on architecture. Contributions. We formalized pragmatic utility as signed change in answer probability, identified the "direction problem" where KL magnitude alone cannot distinguish helpful vs. harmful passages, and provided theoretical guarantees for rejecting task-irrelevant updates (Theorem 3.2). Limitations. ECS requires logit access (limiting use with closed APIs) and exhibits model- dependence and granularity sensitivity. These suggest ECS is best suited for open-weight models with fine-grained context selection. Future Work. Key directions include model-agnostic variants using sampling-based estimation, multi-stage pipelines combining semantic pre-filtering with ECS ranking, and extending to multi-turn reasoning across distributed evidence.

References Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations (ICLR), 2024. Oral presentation (top 1%). Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020. Sebastian Borgeaud et al. Improving language models by retrieving from trillions of tokens. ICML, 2022. Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 17754–17762, 2024. Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595, 2023. Xin Cheng, Xun Wang, Xingxing Zhang, Tao Wu, Yanan Luo, and Furu Wei. xRAG: Extreme context compression for retrieval-augmented generation with one token. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2405.13792. Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2006. Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In Advances in Neural Information Processing Systems (NeurIPS), volume 35, pp. 16344–16359, 2022. Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 150–163, 2024. arXiv:2309.15217. Kawin Ethayarajh et al. Understanding dataset difficulty with v-usable information. ICML, 2022. Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Guo, Wenjun Chen, and Kaiwen Liao. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In Proceedings of ACL, pp. 9962–9977, 2024. Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024. Kelvin Guu et al. Realm: Retrieval-augmented language model pre-training. ICML, 2020. Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Squeezed attention: Accelerating long context length LLM inference. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024. arXiv:2411.09688. Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Acceler- ating and enhancing LLMs in long context scenarios via prompt compression. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024. Vladimir Karpukhin et al. Dense passage retrieval for open-domain question answering. EMNLP, 2020. Woosuk Kwon et al. Efficient memory management for large language model serving with pagedat- tention. SOSP, 2023. Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS, 2020. Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024. Yi Liu et al. RECALL: A benchmark for LLMs robustness against external counterfactual knowledge. arXiv preprint arXiv:2311.08147, 2023. Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 13851–13868, 2024.

Xiurui Pan, Endian Li, Qiao Li, Liang-Chieh Lu, Shi-Xin Shen, Yan Zhuang, Kang Li, Chao Wang, and Shengjie Zhang. InstAttention: In-storage attention offloading for cost-effective long-context LLM inference. In IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2025. doi: 10.1109/HPCA61900.2025.00113. Joon Sung Park et al. Generative agents: Interactive simulacra of human behavior. UIST, 2023. Bowen Peng, Jeffrey Quesnelle, Honglu Fan,

and Enrico Shao. Yarn: Efficient context window extension of large language models. arXiv preprint arXiv:2309.00071, 2023. Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT- networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP), pp. 3982–3992, 2019. Noah Shinn et al. Reflexion: Language agents with verbal reinforcement learning. NeurIPS, 2023. Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–21, 1972. Babak Taati et al. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset. In NeurIPS, 2023. Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 183–196, 2020. Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4396–4406, 2019. Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. In NeurIPS Workshop on Agent Learning in Open-Endedness, 2023. Spotlight. Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS, 2022. Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In Proceedings of the International Conference on Learning Representations (ICLR), 2025. arXiv:2410.10813. Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In ICLR, 2024. Yilun Xu et al. A theory of usable information under computational constraints. ICLR, 2020. An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. Shunyu Yao et al. React: Synergizing reasoning and acting in language models. ICLR, 2023. Zeyu Zhang, Xiaohe Bo Zhang, Chao Chen, et al. A survey on the memory mechanism of large language model based agents. arXiv preprint arXiv:2404.13501, 2024. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In NeurIPS, 2023.

Table 2: Hyperparameter settings Parameter Value Trajectory horizon (T) 8 Information threshold ( ) 0.05 Length penalty () 0.002 Top-k tokens for KL 50 Epsilon smoothing 1010 ACE similarity threshold 0.85 Table 3: Dataset statistics Statistic LongMemEval LoCoMo

Total samples 500 200 Samples used 100 200 Candidates/question $\sim$ 50sessions$ $\sim$ $400turnsTokens/candidate$ $\sim$ $6K$ $\sim$ $50–200Evidenceitems/question1–3sessions2–5turnsAProofsA.1ProofofTheorem3.2Proof.Letubeatask−irrelevantupdatesuchthatuq|C.WewishtoshowthatDKL(PCuPC).Bythedefinitionofconditionalindependence : P(y|Cu, q) = P(y|C, q)(5)forallyV.ThisimpliesPCu = PC, andthus : DKL(PCuPC) = 0(6)Inpractice, finitemodelcapacityintroducessmalldeviations, yieldingDKLforproportionaltomodelapproximati SameQuestions : Bothmethodsevaluatethesamequestionsfrom LongMemEvalandLoCoMoSameContexts : Identicalcandidatesessions/turnsarescoredbybothmethodsSameGroundTruth : Evidencelabelsareheldconstant Thekeydifferenceisthescoringfunction : ACE(TF − IDF) : Scorespassagesbylexicaloverlapwiththequery ECS(LLM − Judge) : ScorespassagesbywhethertheyhelpanswercorrectlyCHyperparameterSettingsThresholdSelection. FortheLLM − as − Judgeoperationalization, weusebinaryclassification : apassageisconsideredhelpfulifthejudgerespondsaffirmatively. Nothresholdtuningisrequiredforthisapproach.C.1 levelselectiontaskwheretheagentmustidentifywhichconversationsessions(outof$ $\sim$ $50)containinformationneededtoanswerthequestion. Eachsessioncontainsapproximately6Ktokensofconversation levelevidenceselectionwheretheagentmustidentifyspecificdialogueturnsthatserveasevidenceforansweringquestio $400turnsacrossmultiplesessions).

Table 4: Detailed experimental results (F1 scores) Dataset Method Qwen2.5-7B Llama-3.1-8B

LongMemEval TF-IDF 0.649 0.649 (Session) Dense 0.591 0.591 ECS 0.152 0.139 LoCoMo TF-IDF 0.154 0.154 (Turn) ECS 0.020 0.265 Table 5: Latency breakdown per candidate (V100 32GB) Component Transformers vLLM Base logprobs 50ms 5ms Candidate logprobs 50ms 15ms KL computation 1ms 1ms Total (cold) 101ms 21ms Total (cached) 51ms 16ms D Detailed Results D.1 Per-Benchmark Performance Table 4 shows the full experimental results with F1 scores for each method

across both datasets and models. D.2 Granularity Analysis The results reveal a clear interaction between context granularity and method effectiveness: • Coarse granularity (LongMemEval): Semantic methods dominate. TF-IDF achieves F1=0.649 vs. ECS F1$\sim$ $0.15. $At session-level($ \sim $6K tokens), noise overwhelms the pragmatic signal. Fine granularity (LoCoMo)$: $ECS excels. With Llama-3.1-8B, ECS achieves F1 = 0.265 vs. TF-IDF F1 = 0.154 - -a 72% relative improvement. D.3 Model Dependence Analysis The stark difference between Qwen(F1 = 0.020) and Llama(F1 = 0.265) on LoCoMo suggests that logprob informativeness varies significantly across model arch$ $Models : Qwen/Qwen2.5-7B-Instruct, meta-llama/Llama-3.1-8B-InstructPrecision : FP16(half precision) Max context length : 8,192 tokens Hardware : NVIDIA V100 32GB Logprobs : Top-20 tokens per position Random seed : 42(for reproducibility of random baselines and any stochastic operations) vLLM Deployment(Optional). For production$ $Prefix caching : Enabled(reduces amortized cost to O(|u|)) GPU memory utilization : 0.85 Recommended GPU : NVIDIA A100 40GB or V100 32GB E.2 Latency Breakdown Table 5 shows the latency compon$ $16ms per candidate.$

F Extended Discussion F.1 Computational Considerations A practical concern with ECS is the computational overhead of computing logprobs for each candidate passage. However, modern inference frameworks provide prefix caching (Kwon et al., 2023) that amortizes context encoding across multiple candidates. Once the base context C is encoded, evaluating each candidate u requires only processing the incremental tokens in u, achieving near-constant overhead with respect to context length. In our experiments, ECS scoring adds approximately 50-100ms per candidate on a single A100 GPU for passages of $\sim$ $100 tokens. Combined with semantic pre-$ $filtering that reduces candidates from thousands to tens, the total latency remains practical for interactive applications.$ $as-Judge Alternative to logprob-based scoring is using an LLM-as-Judge approach where a model explicitly evaluates "Does this passage help answer the question?" Such approaches can cap$ $as-Judge reserved for top candidates requiring careful evaluation. F.3 Error Analysis and Failure Modes Beyond them$ $turn reasoning failures noted in the main text, we identify additional patterns where ECS underperforms :$ $Indirect evidence : When a passage provides background knowledge that indirectly supports the answer rather than stati$ $For questions where the model has low confidence across all possible answers, even relevant passages may not induce suf$ $When the correct answer spans multiple tokens, first-token KL-$ $divergence may not capture the full utility. Our trajectory divergence extension partially addresses this, but optimal hori$ $PyTorch 2.0 + Hugging Face Transformers 4.35 + sentence-transformers vLLM 0.5 +$ $(optional, for faster inference with prefix caching)$

REFERENCES

Asai et al.. 2024.

Beltagy et al.. 2020.

Borgeaud et al.. 2022.

Chen et al.. 2023.

Chen et al.. 2024.

Cheng et al.. 2024.

Cover & Thomas. 2006.

Dao et al.. 2022.

Es et al.. 2024.

Ethayarajh et al.. 2022.

Fang et al.. 2024.

Guu et al.. 2020.

Hooper et al.. 2024.

Jiang et al.. 2024.

Karpukhin et al.. 2020.

Kwon et al.. 2023.

Lewis et al.. 2020.

Liu et al.. 2023.

Liu et al.. 2024.

Pan et al.. 2025.

Park et al.. 2023.

Peng et al.. 2023.

Reimers & Gurevych. 2019.

Shinn et al.. 2023.

Spärck Jones. 1972.

Taati et al.. 2023.

Voita & Titov. 2020.

Wang et al.. 2023.

Wei et al.. 2022.

Xiao et al.. 2024.

Xu et al.. 2020.

Yao et al.. 2023.

Zhang et al.. 2024.

Zheng et al.. 2023.