No More Free Lunch: Corpus Task Complexity Matters as Corpora Grow

Prasann Singhal[α] **Amanda Bertsch**[γβ] **Jacob Steinhardt**[α] **Sewon Min**[αβ]

[α]UC Berkeley [β]Allen Institute for AI [γ]Carnegie Mellon University

prasann@berkeley.edu

Code / Data PrasannS/corpustaskcomplexity

Abstract

Given a large corpus, the questions one might ask can vary—from "When was the first human heart transplant?" to "What are all the contradictory claims in this literature?"—but what makes some questions more challenging than others? In this work, we define a notion of **Corpus Task Complexity (CTC)** that characterizes tasks by how their difficulty grows with corpus size; for instance, a retrieval query only requires a single linear pass over a corpus, while finding contradictions requires checking a quadratically growing set of claim pairs. Observing that prior work has largely only studied tasks whose difficulty grows linearly with corpus size, which we call *low* CTC tasks, we introduce 10 new tasks belonging to a class of *high* CTC whose difficulty grows quadratically or more in corpus size. We find that high-CTC tasks not only grow much more challenging on average at longer contexts for LCLMs, they reverse many modeling conclusions drawn solely from low-CTC evaluations. For instance, efficient block-sparse and hybrid attention approaches consistently match full attention performance on low-CTC tasks, but degrade much more on high-CTC tasks. Large-corpus high-CTC reasoning thus remains an open challenge as full attention is too costly to scale, motivating future research on these tasks. We release our code, data, and 22-task suite (CTC-BENCH), to facilitate future research in this area.

1 Introduction

Computational tools enable us to conduct search over large digital corpora—from scientific literature to the internet—but is it yet possible to develop systems that could, for example, *find all contradictions* in a large corpus? Given the wealth of valuable information contained in large corpora, researchers have built diverse systems to extract and analyze this information, from work in information retrieval [Craswell et al., 2025, Hou et al., 2025] to summarization [Kociský et al., 2017]. We refer to these collectively as *corpus reasoning* tasks, but they can vary substantially in their complexity over the corpus, ranging from simple factoid queries like "When was Ralph Lauren founded?", to tasks requiring extensive cross-document interaction, such as "Find all contradicting claims in this biology literature" or "What are the outliers in the LLM agent trace corpus?".

Long-context language models (LCLMs) trained to process large inputs (e.g., 32K tokens or more) are promising for corpus tasks—with softmax attention enabling direct retrieval and reasoning across parts of the conditioned corpus. Modern LCLMs achieve strong performance on many existing synthetic long-context benchmarks [Yen et al., 2024, Bai et al., 2024, Chen et al., 2026], even with various efficient approaches used to reduce their computational cost [Yang et al., 2024, Beltagy et al., 2020]. But given the lack of clear characterization of what makes corpus reasoning tasks difficult

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

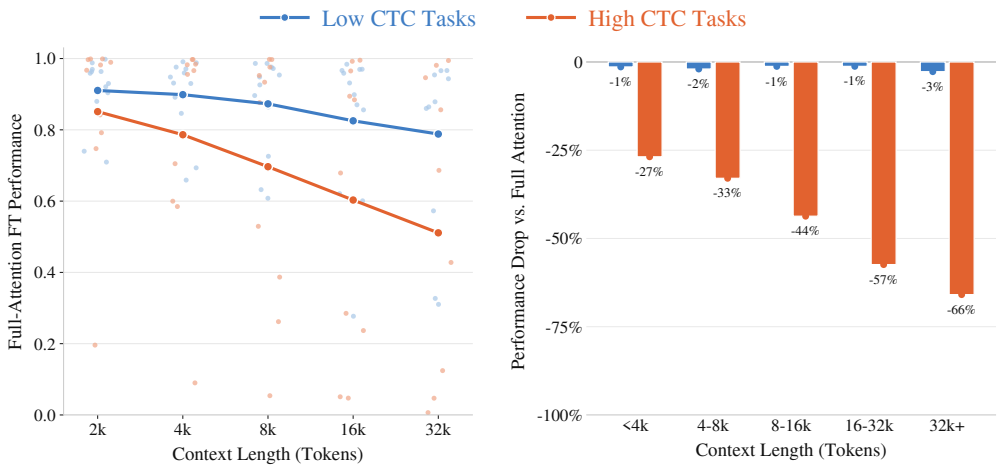

Figure 1: Performance of in-distribution trained Qwen3.5-4B models on low- and high-CTC tasks. *(Left)* Performance degrades faster on high-CTC tasks than on low-CTC tasks as corpus size increases. *(Right)* Block-sparse attention causes almost no performance degradation relative to full attention on low-CTC tasks, but substantial degradation on high-CTC tasks.

on large corpora, it remains unclear whether these successes apply to all corpus reasoning tasks, particularly more complex tasks.

We begin by defining *corpus task complexity* (CTC), which characterizes a task by how the number of operations required to solve it scales with corpus size. For example, the number of local operations required to retrieve a fact usually grows linearly with corpus size, since each document need only be checked once to find where the fact is located. In contrast, a task such as "Find all contradicting claims in the literature" involves comparing all pairs of claims, resulting in quadratic growth. Through this lens, we survey 12 diverse corpus reasoning tasks commonly used to evaluate LCLMs and find them to all grow linearly in difficulty with corpus size (which we call *low-CTC*). We then introduce 10 diverse tasks whose complexity grows quadratically or faster (*high-CTC* tasks), ranging from identifying all contradictory claims in a corpus to finding documents with the rarest topics.

Using this benchmark suite, CTC-BENCH, we evaluate a range of LCLMs across corpus sizes from 2K to 32K tokens. We first find that with increasing corpus size, performance degrades faster on average on high-CTC tasks than on low-CTC tasks (Figure 1(a)). More importantly, however, **we find that high-CTC task results frequently reverse modeling conclusions** drawn solely from low-CTC evaluations. For example, block-sparse attention—an efficient attention variant often used for corpus reasoning [Xiao et al., 2025]—shows minimal degradation on low-CTC tasks, yet on high-CTC tasks it sees large performance drops relative to full attention (Figure 1(b)). Similar discrepancies occur for other modeling decisions: hybrid architectures that alternate linear and full attention exhibit larger performance gaps on high-CTC tasks; and length generalization—fine-tuning a long-context model on shorter contexts than the intended context length at test time [Gao et al., 2024]—can work for low-CTC tasks but breaks down on high-CTC tasks.

To summarize, our results show that evaluations including high-CTC tasks reveal hidden costs to many common modeling choices, and that there is no free lunch: full attention becomes quadratically intractable on large corpora, but efficient attention mechanisms that appear lossless under existing low-CTC evaluations can lead to substantial degradation on high-CTC tasks. Solving high-CTC tasks on large corpora is a challenging long-term problem, and we encourage more research into scalable architectures that can handle both low *and* high-CTC tasks. To support this effort, we release CTC-BENCH, along with our code and data.

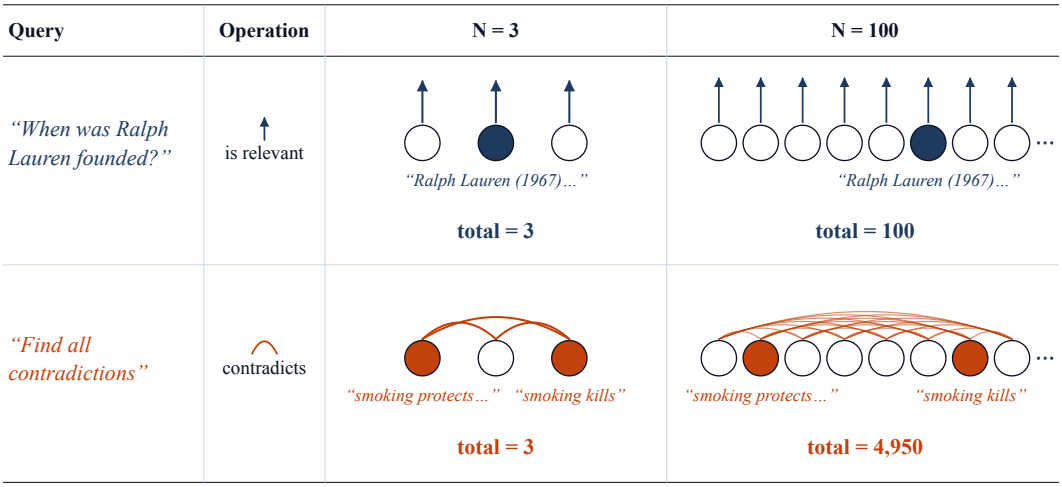

Figure 2: Circles represent documents, with shaded circles for the gold documents. As corpus size $N = |C|$ grows from 3 documents to 100 documents, a contradiction search task may in worst-case require many more oracle operations (4950) than factoid retrieval (100).

2 Corpus Task Complexity & CTC-BENCH

What most strongly determines corpus task difficulty? We propose *Corpus Task Complexity (CTC)*, a definition of task difficulty based on how a task's difficulty scales with corpus size (§2.1), which we walk through with two tasks (shown in Figure 2). We find that most existing benchmarks are low CTC—their complexity grows linearly with corpus size (§2.2)—then introduce a new class of high CTC tasks, whose complexity grows quadratically or faster (§2.3).

2.1 Corpus Task Complexity

Define a corpus C to be an unstructured set of N documents $\{d_1, \ldots, d_N\}$, where each d_i is a sequence of tokens. A corpus reasoning task T is a set of same-type corpus-question-answer samples $(C_i, q_i, a_i) \in T$. For instance a factoid retrieval task sample could be (C_i =Wikipedia, q_i ="*When was Ralph Lauren founded?*", a_i="*1967*"), and a contradiction task sample could be (C_i =CommonCrawl, q_i = "*Find all contradictions*", a_i =list of contradicting pairs within C_i).

To understanding corpus reasoning tasks, we define a notion of complexity called *corpus task complexity (CTC)*. Imagine that we can call an *oracle* (e.g. an LM-judge) that can answer questions over small contexts with O(1)-length in corpus size. CTC is the asymptotic growth in oracle calls needed to solve tasks as a function of corpus size N.

Specifically, let A_T be an algorithm composed of oracle calls that solves *all* $(C_i, q_i, a_i) \in T$, where $f_{A_T}(C_i, q_i)$ indicates oracle call count, and we constrain A_T to be the algorithm with lowest possible call count across samples. Task T is $O_T(g(N))$ if there exists some constant α such that $f_{A_T}(C_i, q_i) \leq \alpha g(N)$ for all $(C_i, q_i, a_i) \in T$. In practice, as proving the optimal A_T can be challenging, we classify CTC with *candidate* best algorithms for tasks given unstructured corpora.

Example 1: (Figure 2, blue) To solve factoid retrieval queries like q = *"When was Ralph Lauren founded?"* with oracle calls, A_T may be to run $\text{oracle}(\text{isrelevant}, d_i, q)$ individually on all $d_i \in C$. Since $f_{A_T}(C_i, q_i)$ is just N, and less operations aren't possible, this task is $O_T(N)$. [1]

Example 2: (Figure 2, orange) To *"Find all contradicting claim pairs."* with oracle calls, A_T may be to run $\text{oracle}(\text{contradicts}, d_i, d_j)$ on all viable document pairs. While priors or corpus structure could lower this, exhaustively checking *all* possible pairs will generally require quadratic oracle calls in corpus size (e.g. 4,950 calls for 100 documents)—making the task $O_T(N^2)$.

[1] CTC should *not* be confused with individual online or offline cost, which combine to become overall cost. For example, nearest neighbors search often has $O(log(N))$ *query-time* complexity, but still requires $O(Nlog(N))$ *build-time* cost.

$O(N)$: Low Complexity Tasks		Higher-complexity tasks	
Dataset	**Description**	**Dataset**	**Description**
NIAH-contra	Find the contradicting claim	Contradiction	Find all contradicting claim pairs
SciFact	Retrieve evidence for a scientific claim	X-Absence	Find unmatched docs across 2 shuffled, near-identical corpora
FiQA	Retrieve relevant financial opinions	QDmatch (HPQA)	Match questions to two documents
MS MARCO	Retrieve relevant web passages	QDmatch (NQ)	Match questions to documents
OBLIQ	Retrieve passages for subjective queries	QDmatch (FiQA)	Match financial queries to documents
NQ	Retrieve the answer passage	Outlier (scale M)	Find chunks from the rarest article
HotpotQA	Retrieve two supporting passages	Grouping (OpenAlex)	Group abstracts by a latent topic
MS MARCO rerank	Re-rank passages by relevance	Strmatch	Find pairs sharing a word sequence
OOLONG	Aggregate labels over the corpus	Reorder	Recover the order of shuffled segments
Outlier (Amazon)	Find reviews with the minority attribute	Textgroups	Find passage triples satisfying a target
Outlier (fixed M)	Find items from the rarest article		
Absence	Identify content missing from a copy		

Table 1: **Overview of the 22 CTC-BENCH Tasks.** Linear-time tasks are shown on the left; tasks requiring higher-order comparisons are shown on the right (details, examples in Appendix H)

Since CTC corresponds to minimal computation *necessary* to solve tasks with growing N, we posit that it will correspond to empirical trends for even black-box methods applied to corpus tasks (e.g. LCLMs, dense retrievers). Similar methods may work for tasks within the same CTC, and on larger corpora an $O_T(N^2)$ task may demand methods with greater inference computation.

Not all same-CTC tasks are identical—retrieval queries with oracle calls involving exact word matches are harder in a constant sense than ones involving abstract relationships, and such differences may even multiply at higher CTC. But by defining task CTC we can better explain what underlies a task's difficulty, and make better modeling decisions to solve it.

2.2 Existing Tasks Through the Lens of CTC

Which CTC categories do existing corpus reasoning tasks belong to? To answer this, we examine a range of 12 well-established existing tasks. We introduce tasks in the order shown in Table 1 (left):

Information retrieval is arguably the best-studied corpus reasoning task. Within it we examine a synthetic recall *NIAH-contra* task, standard *SciFact, FiQA, and MS-MARCO* tasks taken from BEIR [Thakur et al., 2021], and the twitter set of the modern *OBLIQ* [Tchuindjo et al., 2026] benchmark. While ranging in difficulty, as discussed in our earlier factoid example these tasks are $O_T(N)$.

We include retrieval variants of two factoid QA datasets: NaturalQuestions (*NQ*) [Kwiatkowski et al., 2019], a single-hop task, and HotpotQA [Yang et al., 2018], a multi-hop task. Both call $\mathrm{oracle}(\mathrm{hasanswer}, q, d_i)$. Since, even in the multi-hop case requiring multiple retrieval steps, only a constant number of passes are required—one retrieval step to find the first hop, and one retrieval step to find the second hop—we classify these also as $O_T(N)$. Likewise, MSMARCO rerank, which involves finding top-10 most relevant documents to a query, given a small k, is at most a constant number of retrieval passes, and thus is also $O_T(N)$.

We include several corpus aggregation tasks. *OOLONG*[Bertsch et al., 2025] involves distributional or counting questions about class labels of documents. *Outlier (Amazon)* involves finding the least common star or product category in a set of reviews. Given a fixed set of class labels, a single pass of $\mathrm{oracle}(\mathrm{classify}, d_i)$ (e.g. 2-stars, 4-stars) calls with lightweight aggregation is typically sufficient for these tasks, making them $O_T(N)$. Lastly, we include a 2-corpus *Absence* task based on AbsenceBench [Fu et al., 2025]—searching for deletions in near-identical Project Gutenberg passages. Since positions are aligned, the oracle is $\mathrm{oracle}(\mathrm{matches}, d_i, d_{i-N})$, where N is individual corpus size. Only a single pass over the second corpus is needed here, so this is *also* $O_T(N)$.

The above tasks constitute a recent, diverse, and representative sample of the most studied IR and LCLM benchmark tasks, yet **they all fall into** $O_T(N)$.

2.3 CTC-BENCH and High CTC Tasks

Our definition of CTC suggests that a variety of useful tasks fall into higher complexity classes, yet such tasks, while explored at small scale (e.g. event re-ordering Chambers et al. [2014]), are understudied for larger corpora, especially in the context of LCLMs. We thus, to enable our analysis, construct a diverse suite of 10 high-CTC tasks (Table 1, right). We call these—along with the 12 low-CTC tasks—**CTC-BENCH** [2] :

(a) *Contradiction* q_i *="Find all the contradicting claim pairs in the corpus."*, a_i is a list of 3 contradicting claim pair IDs. It uses LLM-generated contradictions to pubmed abstract claims [Jin et al., 2019], and is as discussed earlier $O_T(N^2)$ (Section 2.1) .

(b) *X-Absence* q_i *="Given near-identical corpus A and B with shuffled chunks, find chunks that are only in one corpus"*, a_i is a list of chunks only in one of either A or B. All d_i are chunks of a Project Gutenberg passage. C_A and C_B (N chunks each) are shuffled to prevent positional bias (unlike Absence which is ordered), so total $\text{oracle}(\text{matches}, d_i, d_j)$ calls are $O_T(N^2)$.

(c-e) *(QDmatch)-FiQA, NQ, HotpotQA* q_i*="Given corpus A with single-sentence questions and B with documents, find (question, document) pairs where documents answer questions"*, a_i is a list of 3 matching pair IDs. We use 3 such settings, using retrieval questions and documents from the FiQA, NQ, and HotpotQA tasks to construct this task. Since C_A and C_B are size N each, cross checking all pairs with $\text{oracle}(\text{isrelevant}, d_i, d_j)$ is $O_T(N^2)$.

(f) *Outlier (wiki)* q_i*="Given the corpus with chunks discussing different topics, identify chunks belonging to the least common topic"*, a_i is a list of minority topic chunk IDs. We source chunks (d_i) from M Wikipedia pages, where the "topic" is the page source (e.g. "Horses"). Unlike OOLONG, which has fixed classes, page topics are diverse and don't repeat across corpora. Given $\text{oracle}(\text{samecategory}, d_i, d_j)$, since each new document only needs to check one document per category, this is $O_T(NM)$. To test this we include an Outlier (fix M) setting where M stays constant ($O_T(N)$) and Outlier (scale M) where M grows proportional to N ($O_T(N^2)$).

(g) *Grouping*, q_i*="Group the following scientific abstracts into* k *groups"*, a_i is a dictionary which must have k keys and put abstracts in the same topic group together (note k increases naturally with N). Abstracts (d_i) and true topic labels (e.g. Biology, Plant Biology) are mined from OpenAlex [Priem et al., 2022]. This has the the same structure as Outlier (wiki) and is $O_T(NM)$.

(h) *Strmatch*, q_i*="Find word sequences with at least k words in common"*, a_i= 3 pair IDs that match. d_i are random noun sequences. Calling $\text{oracle}(\text{hascommonk}, d_i, d_j)$, the task requires searching over $O_T(N^2)$ pairs. This task serves as a synthetic control for other pairwise tasks.

(i) *Reorder* (Bai et al. [2024]), q_i*="Given the passage (chunks have been randomly shuffled), output the original ordering."*, a_i is an ID list of length N with true ordering, and d_i are sentences from a Project Gutenberg passage. If we use $\text{oracle}(\text{directlyneighbors}, d_i, d_j)$, this requires brute-force checking all pairs to recover the order ($O_T(N^2)$).

(j) *Textgroups*, q_i*="Find groups of 3 passages whose adjective count add ups to t."* given Project Gutenberg chunks, where property and count t vary (e.g. 67 nouns). With calls to $\text{oracle}(\text{propertysum}, d_i, d_j, d_k)$ searching over all triples, we include it as an $O_T(N^3)$ exemplar.

3 Role of Task Complexity on Growing Corpora

Having established CTC and CTC-BENCH, we now evaluate a range of models to understand how well they perform on these tasks. In particular, we focus on performance as a function of corpus size: *How challenging do high-CTC tasks become as corpus size grows?*

Experimental Setup We train Qwen3.5-4B individually for all 22 CTC-BENCH tasks, ranging from 2k to 32k context length in training and evaluation. For all tasks except OBLIQ (where we use synthetic data, Appendix C), we use existing or generated in-domain fine-tuning sets to isolate architectural limits independent of data and generalization.

[2] We often include synthetic / semi-synthetic tasks for controllability and cleaner analysis, though we design all CTC-BENCH tasks to structurally resemble realistic tasks (see Appendix B)

We always train for 1 epoch on 20k datapoints per task (except for OBLIQ and SciFact), with evenly split examples across context lengths (matching the corpus sizes we evaluate on). We train with full fine-tuning, with a learning rate of 5×10^{-5}. All tasks use this prompt format:

```
[Initial] 1:[Doc 1 content] ...N:[Doc N content] [Query] [Answer]
```

Result: High-CTC Task Performance Drops More Rapidly Than Low-CTC. We plot in Figure 1 (left) *averages across tasks* of full attention performance for all High-CTC (orange) and Low-CTC (blue) tasks across CTC-BENCH (task-individual plots in Figure 3). Average performance drops $0.910 \rightarrow 0.788$ (13%) from 2k to 32k on low-CTC (10 out of 12 tasks degrade less than 0.1 in performance) but on high-CTC tasks average performance drops $0.851 \rightarrow 0.511$ (40%) from 2k to 32k (7 of 10 tasks drop over 0.1 in performance). In aggregate, even with in-distribution training, high complexity task difficulty grows much faster on larger corpora.

4 Rethinking Free Lunches With CTC

In the previous section, we show that high-CTC tasks often degrade at larger corpus sizes much faster than more commonly studied low-CTC tasks—suggesting that even costly full attention models are insufficient at scale for high-CTC tasks. But how does this relate to the large body of prior LCLM work that has predominantly focused on *reducing* LCLM computation? We revisit 3 common modern LCLM choices from the lens of CTC: block-sparse attention (§4.1), hybrid architectures (§4.2), and length generalization from short train contexts (§4.3).

4.1 Full vs Block-Sparse Attention

In a regular Transformer model (full attention), pre-filling a corpus has quadratic computational complexity. Motivated by the fact that in corpus reasoning, the input is a collection of independent documents, much prior work has tried [Beltagy et al., 2020, Zaheer et al., 2020, Gollapudi et al., 2026, Xiao et al., 2025] block-sparse attention: pre-filling each document independently so that they attend to tokens within the same document only, with only a small number of query and answer tokens attending globally. These approaches have been shown effective, though often limited to evaluation on $O_T(N)$ settings like QA, retrieval, and in-context learning.

We investigate several decisions reducing computation from full attention, but what if we instead apply *more* computation to cheap methods? Our block-sparse attention implementation follows that of [Xiao et al., 2025] [3], but we additionally propose a new *mask-mixing* technique: during block-sparse attention training, with probability p a training batch uses the full-attention mask instead of the block-sparse one, where $p = 0$ indicates original block-sparse attention, and $p = 1$ indicates full attention (we anneal on a curriculum from $p = 0.8$ to $p = 0$ during training). During evaluation, block-sparse attention is used for all inputs.

Mask-Mixing Frequently Improves Block-Sparse. We evaluate mask-mixing on a 10 task subset that we call CTC-BENCH-10, that we select to cover diverse task structures and difficulties across CTC classes in CTC-BENCH (results in Figure 5, Left). We find mask-mixing greatly improves block-sparse attention, and *consequently apply it to all block-sparse experiments*. Specifically, while it helps less on retrieval settings like FiQA and NQ, we find that on many High-CTC tasks and tasks requiring aggregation across context like OOLONG, mask-mixing improves block-sparse performance—potentially by distilling representations for tasks from the more computationally expressive full attention.

Full vs Block-Sparse Gap Grows Faster in High-CTC. Figure 3 shows degradation and comparisons between full and block-sparse attention for the full CTC-BENCH. On low-CTC tasks, the gap between full and block-sparse attention is negligible *for all tasks*—reproducing much prior work [Beltagy et al., 2020, Gollapudi et al., 2026], we find that block-sparse gives a free lunch of comparable performance (sometimes even better) while being much cheaper. However, across all high-CTC tasks, choosing block-sparse attention over full attention leads to worse performance, despite our own improvements through mask-mixing. Unless the task is so difficult that both approach zero (e.g. Textgroups, Reorder), this gap monotonically grows at longer contexts. Averaged across tasks, the performance degradation from using block-sparse on high-CTC is -26.9% (2k), -33.0%

[3] Qwen3.5 models use 75% gated delta-net layers (already linear), so we only apply block-sparse to remaining 25%

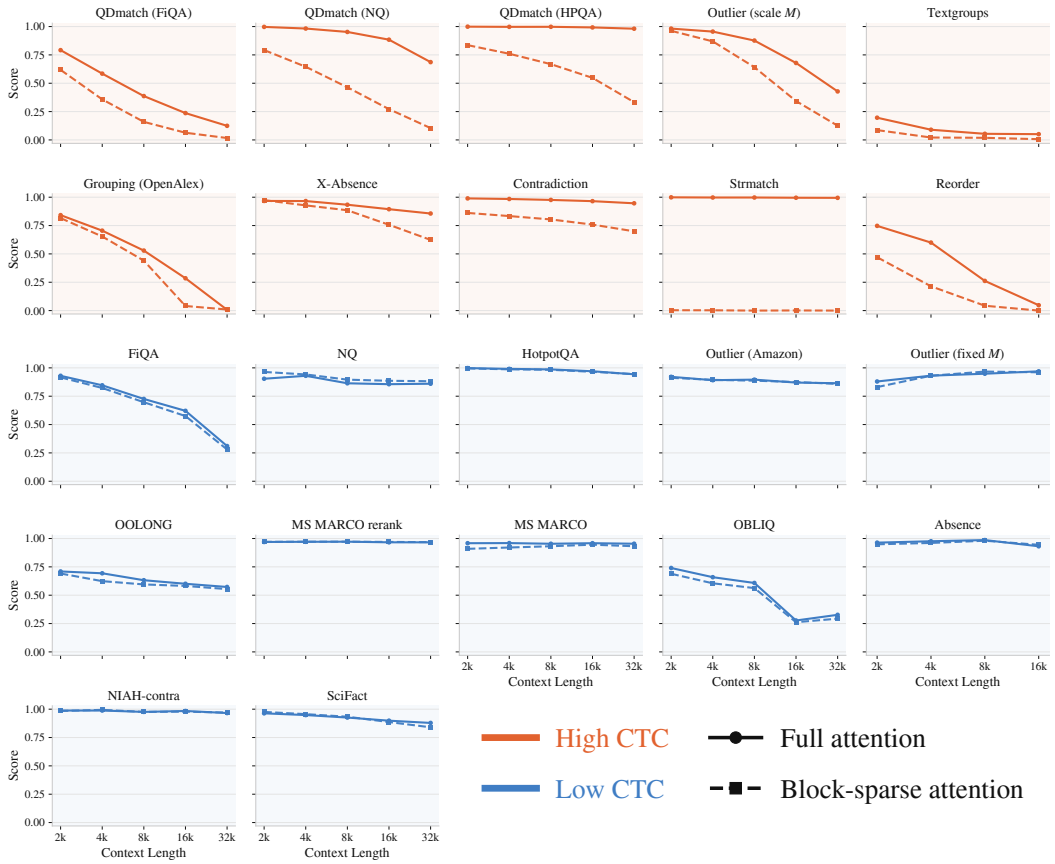

Figure 3: *Scaling of full vs. block-sparse attention (mask mix)* We show block-sparse vs full attention performance at different contexts on our full task suite. While both methods perform near-identically on low-CTC, high-CTC tasks correspond to growing gaps between them on larger corpora.

(4k), -43.7% (8k), -57.4% (16k) and -65.9% (32k). This supports a central theme—decisions that did not matter for low-CTC matter greatly on high complexity tasks.

4.2 Full vs Hybrid

Beyond block-sparse attention, alternative architectures such as state-space-models (SSMs) [Katharopoulos et al., 2020, Gu and Dao, 2023] and adjacent approaches like *gated delta net (GDN)* [Yang et al., 2024] have also enabled attention to reduce computational complexity. Specifically, these methods replace all-to-all token interactions with a recurrent state update, giving linear computation and constant memory. While this has shown promise, pure-GDN has, even at low-CTC, proven insufficient—language models like Qwen3.5 [Qwen Team, 2026] have mixed GDN with full attention rather than replacing it.

Strong performance of Qwen3.5 on long-context benchmarks Bai et al. [2024]) and prior comparisons of full vs hybrid architectures [Merrill et al., 2026] have suggested hybrid architectures to be comparable or even better than full attention, but does this free lunch apply to high-CTC tasks? To do a controlled test of this question, we compare OLMo-3-7B [Ettinger et al., 2025] vs OLMo-3-7B-Hybrid [Merrill et al., 2026], the closest open-weight comparison between hybrid and full architectures. Note that since Olmo3 by default uses sliding window attention, we adapt it to a full-attention version with a 100M-token continued-pretraining phase on Dolma3-Longmino data (0.1% of its mid-training budget). OLMo-3-7B-Hybrid is already mid-trained natively on this data.

Hybrid Underperforms Full Attention on High-CTC Tasks. We plot individual task results on the 10 CTC-BENCH-10 tasks in Figure 4. Similar to block-sparse attention, hybrid attention often leads to no degradation compared to full attention on low-CTC tasks (in fact doing better at 32k),

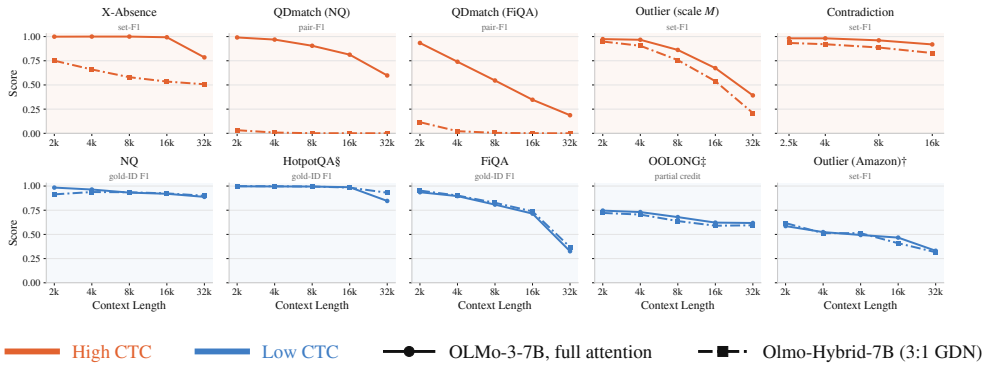

Figure 4: *Hybrid vs Full Models* We plot CTC-BENCH-10 results for a full-attention (solid) version of OLMo-3-7B vs OLMo-3-Hybrid (dot-dashed). On low-CTC tasks both perform similarly, but on high-CTC tasks we observe consistent performance gaps.

Low CTC		High CTC	
Task	Gain	Task	Gain
NQ	+0.002	QDmatch (FiQA)	+0.031
HotpotQA	+0.004	QDmatch (NQ)	+0.283
FiQA	+0.011	Outlier (scale-M)	+0.505
OOLONG	+0.581	XAbsence	+0.646
Outlier (Amazon)	+0.773	Contradiction	+0.773
Mean	+0.274	*Mean*	+0.448

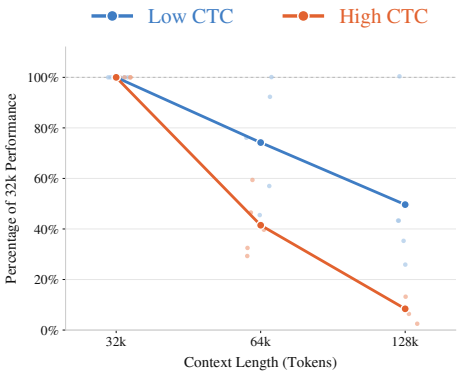

Figure 5: **(Left)** *Gain of curriculum mask-mixing over pure block-sparse training (no mixing)*, averaged across 2k–32k contexts on CTC-BENCH-10. Mask-mixing consistently improves block-sparse, especially on High-CTC tasks and tasks more dispersed across contexts (such as OOLONG). **(Right)** *Length generalization (full attention).* For 64k and 128k context eval sets (trained only to 32k), we plot percentage of 32k score. High-CTC tasks are harder for length generalization.

but leads to large degradation on high-CTC. Note that these numbers are affected by OLMo models being generally weaker than Qwen3.5, and while overall trends are consistent, the gap varies by task, e.g. QDmatch (NQ) vs Outlier (wiki, Scale M).

4.3 Length Generalization

Our work shows that even up to the relatively small scale of 32k tokens, high-CTC tasks already degrade greatly—scaling data sufficiently, especially for larger corpora, becomes increasingly important. But LCLM training is costly. Given these prohibitive costs, much work has sought to use short-context data to generalize to longer contexts (training only on short versions of tasks or short instruction-tuning data). Such work has successfully shown generalization on a variety of tasks past lengths used in fine-tuning [Gao et al., 2024, Peng et al., 2023, Mehta et al., 2026]. However, these results have predominantly been demonstrated only on low-CTC settings.

High-CTC Tasks May Generalize Worse to Longer Contexts. We evaluate 2k-32k-trained models at 64k and 128k on CTC-BENCH-10. Note this does not require context extension as Qwen3.5-4B's native context is $\tilde{2}$56K tokens. We plot results in Figure 5 (right) with full attention, showing percentage of 32k performance retained on the y-axis to increase comparability. At 128k context low-CTC ranges from 28%-100% while high-CTC ranges from 4%-14%—all high-CTC points fall below low-CTC points. We find these trends at only 128k tokens, and it's likely that this issue may

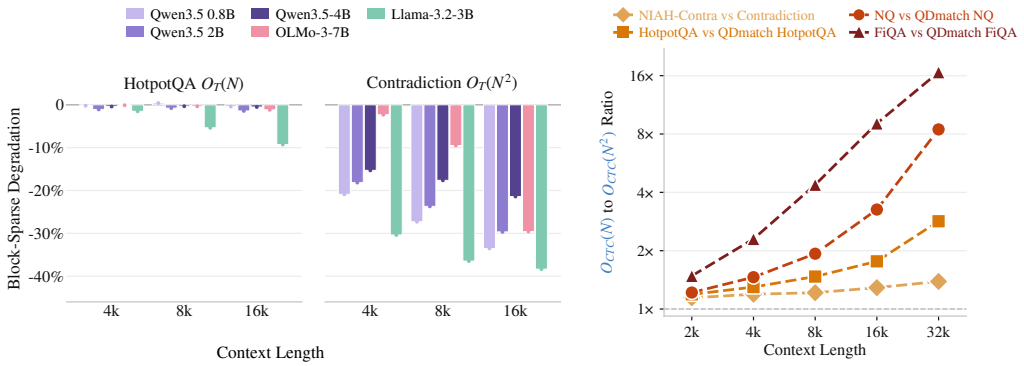

Figure 6: **(Left)** *Model scale/family* Percentage of full-attention performance lost to block-sparse attention at 4k, 8k and 16k context, for three Qwen3.5 scales and two other model families, on a low-CTC (HotpotQA, $O_T(N)$) and high-CTC (Contradiction, $O_T(N^2)$) task. The gap is minimal on low-CTC and large and growing on high-CTC; within the Qwen3.5 family (shades of purple) smaller models degrade faster. **(Right)** *Oracle Difficulty* For 4 task pairs with matched oracle operations (block-sparse attention), ratio of $O_T(N)$ to $O_T(N^2)$ task performance on the same corpus (pairs named in the legend). **Tasks with harder base operations (darker) see the ratio grow faster.**

become even more pronounced at larger corpus scales (e.g. 1 million+ tokens). Taken together with our prior results, this poses a conundrum: full attention is too costly to scale to large corpora, but many modeling decisions that make LCLMs tractable for larger corpora cause degradation on high-CTC tasks.

5 Analyses: Model Scale, Operation Difficulty

We so far show that high-CTC often translates to tasks becoming much harder on larger corpora and how various "free lunches" that have been demonstrated on common tasks don't apply on high-CTC tasks. In this final section, we analyze how factors beyond complexity, both modeling (model family, model scale, §5.1) and task-specific (oracle difficulty, §5.2) interplay with our previous findings.

5.1 Role of Model Family and Scale

Are our findings robust to different model scales and families? To answer this question, we sample two tasks: HotpotQA retrieval, one of the most commonly evaluated low-CTC tasks in CTC-BENCH, and Contradiction, a representative high-CTC task. We train with three model families (OLMo-3-7B, Qwen3.5-4B, and Llama3.2-3B) [Team, 2024, Ettinger et al., 2025], as well as different model scales within Qwen3.5. We compare full vs block-sparse attention performance to see whether the trends align with our earlier finding.

Block-Sparse Result Holds Across Model Families. We report the performance gap between full and block-sparse attention across three sizes of Qwen 3.5, Llama, and OLMo in Figure 6 (Left); individual results are reported in Figure 8 of Appendix E. Across all families and scales, block-sparse attention causes little degradation on HotpotQA but substantial degradation on Contradiction. Comparing degrees of degration across three model sizes of Qwen-3.5 (0.8B, 2B, 4B) additionally reveal the pattern that smaller models degrades faster on larger input corpora. This confirms our earlier finding that replacing full with block-sparse attention has a larger impact on high-CTC than low-CTC tasks.

5.2 Oracle Operation Difficulty

CTC focuses on how the number of oracle calls required to solve a task scales with corpus size, rather than the difficulty of each oracle call itself. A natural question to ask is: How does oracle operation difficulty affect model performance on low- and high-CTC tasks?

To study this, we consider four $O_T(N)$ tasks, Contra-NIAH-Contra, HotpotQA, NQ, and FiQA, and rank them by relative difficulty, measured empirically by Qwen3.5-4B performance averaged across all evaluated context lengths. We pair each task with its corresponding $O_T(N^2)$ tasks—Contradiction, QDMatch HotpotQA, QDMatch NQ, QDMatch FiQA—and measure how much harder the $O_T(N^2)$ version is relative to its matched $O_T(N)$ task, i.e., the ratio of their performance at aech context length, under block-sparse attention.

CTC Magnifies Difficulty Gaps Between Tasks. Figure 6 (Right) shows that the performance gap between each $O_T(N^2)$ task and its matching $O_T(N)$ tasks grows with context length. Importantly, the ordering of these gaps follows the ordering of oracle difficulty: tasks with harder oracle operations exhibit larger gaps as context length increases. This suggests that oracle difficulty and CTC interact systematically, and that relative degradation between tasks at shorter contexts may provide bounds on, or even predict degradation at longer contexts and higher CTC classes.

6 Related Work

Prior Corpus Reasoning Tasks Corpus reasoning tasks (including examples with possibly high-CTC) have been studied for decades in computer science, from various low-CTC tasks in our suite (e.g. IR Craswell et al. [2025]), to others like event ordering [Chambers et al., 2014]. Fields like Exploratory Data Analysis [Tukey, 1962] (which studies visualization and understanding of large data), though often focused on structured data, have explored engineering specialized pipelines to understand unstructured data [Shankar et al., 2024]. Work in library sciences has also developed theories of interrelatedness between documents, most famously in the taxonomy of Tillett [1991]. More recently Zhang et al. [2025]—which informally mentions a notion of task complexity related to CTC—includes an exhaustive pair generation task they find to be more challenging at scale for closed-source LLMs. Overall, while we assemble a large and new set of high-CTC tasks, we are not the first to propose tasks classifiable as high-CTC. That said, beyond existing notions [Goldman et al., 2024], we develop a new unified understanding of difficulty within CTC classes. To our knowledge this work is the first to identify LCLMs as uniquely suitable for high CTC tasks, and to connect task complexity to architectural decisions. Through CTC-BENCH, our work also provides the most comprehensive testbed for how high-CTC task difficulty grows at scale.

Scalable Corpus Methods Work applying language models to corpora has taken several directions. One path is to develop agentic scaffolds that combine efficient tools like dense retrievers Khattab and Zaharia [2020], Karpukhin et al. [2020] with LLMs. These are often hampered by fundamental challenges in retrieval [Weller et al., 2025, Su et al., 2024, Wei et al., 2025], but as shown by Zhang et al. [2025] can be promising given the right scaffold and sufficient inference compute. Note that the slowness of auto-regression (and the high per-query online cost, Zhang et al. [2025]) make agentic systems intractable for high-complexity tasks—finding subtle contradictions via enumeration across N documents may require generating on the order of N^2 tokens per query. Consequently, a complementary line of work—which we focus on in this work—seeks to develop architectures capable of ingesting full corpora end-to-end, often by making attention more efficient [Acharya et al., 2024, Beltagy et al., 2020, Ivgi et al., 2022, Lu et al., 2024, Xiao et al., 2025, Izacard and Grave, 2020, Zaheer et al., 2020, Katharopoulos et al., 2020, Yang et al., 2024]. Context extension Peng et al. [2023], and length generalization Mehta et al. [2026] also fall into these efforts. Our work helps develop a unified understanding of what these can / cannot achieve as a function of their computation. Importantly, our findings qualify this line of work, and reveal limitations obscured by the fact that these methods have primarily been developed on low-CTC tasks.

7 Conclusion

Our work focuses on clearly defining the challenges around high-CTC settings, and identifying the relationship between architecture design and CTC. While helping explain the success of cheaper linear cost methods, our investigation in particular encourages the community to avoid over-reliance on low-CTC evaluations and pay greater attention to high CTC tasks, which represent an interesting and impactful long-term open problem.

Acknowledgment

We thank Ryan Wang, Karim Abdel Sadek, Jongho Park, Diane Tchuindjo, Omar Khattab, Suhas Kotha, Rhys Gould, Yaowen Ye, Yichuan Wang and other members of Jacob Steinhardt, Sewon Min, and Berkeley AI Research groups for discussion and feedback. This research was supported in part by ONR (N00014-26-1-2233), the NVIDIA Academic Grant Program, and gifts from Ai2 and Apple. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Numbers DGE2146752, DGE2637800, DGE2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project was also supported generously by compute from VESSL AI.

References

Shantanu Acharya, Fei Jia, and Boris Ginsburg. Star attention: Efficient llm inference over long sequences. *ArXiv*, abs/2411.17116, 2024. URL https://api.semanticscholar.org/CorpusID:274280741.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *ArXiv*, abs/2412.15204, 2024. URL https://api.semanticscholar.org/CorpusID:274859535.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150, 2020. URL https://api.semanticscholar.org/CorpusID:215737171.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. Oolong: Evaluating long context reasoning and aggregation capabilities. *ArXiv*, abs/2511.02817, 2025. URL https://api.semanticscholar.org/CorpusID:282749185.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2: 273–284, 2014. URL https://api.semanticscholar.org/CorpusID:1564278.

Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qingfang Fu, Debing Zhang, and Songlin Hu. Longbench pro: A more realistic and comprehensive bilingual long-context evaluation benchmark. *ArXiv*, abs/2601.02872, 2026. URL https://api.semanticscholar.org/CorpusID:284512608.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Fernando Campos, and Jimmy J. Lin. Overview of the trec 2021 deep learning track. *ArXiv*, abs/2507.08191, 2025. URL https://api.semanticscholar.org/CorpusID:261242374.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David W. Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Daniel Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee F. Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler C. Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh M. Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James Validad Miranda, Maarten Sap, Malia Morgan, Michaela Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, William Merrill, Yapei Chang, Yuling Gu, Zhi yuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. *ArXiv*, 2025. URL https://api.semanticscholar.org/CorpusID:283908770.

Harvey Yiyun Fu, Aryan Shrivastava, Jared Moore, Peter West, Chenhao Tan, and Ari Holtzman. Absencebench: Language models can't tell what's missing. *ArXiv*, abs/2506.11440, 2025. URL https://api.semanticscholar.org/CorpusID:279391603.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:273098476.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. *ArXiv*, abs/2407.00402, 2024. URL https://api.semanticscholar.org/CorpusID:270870356.

Siddharth Gollapudi, Nilesh Gupta, Prasann Singhal, and Sewon Min. Can language models actually retrieve in-context? drowning in documents at million token scale. *ArXiv*, 2026. URL https://api.semanticscholar.org/CorpusID:289749122.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023. URL https://api.semanticscholar.org/CorpusID:265551773.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn J. Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. Clerc: A dataset for u. s. legal case retrieval and retrieval-augmented analysis generation. In *North American Chapter of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:278664835.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *ArXiv*, abs/2403.03952, 2024. URL https://api.semanticscholar.org/CorpusID:287635730.

Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299, 2022. URL https://api.semanticscholar.org/CorpusID:251224058.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *ArXiv*, abs/2007.01282, 2020. URL https://api.semanticscholar.org/CorpusID:220302360.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43, 2022. URL https://api.semanticscholar.org/CorpusID:251371732.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *ArXiv*, abs/1909.06146, 2019. URL https://api.semanticscholar.org/CorpusID:202572622.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020. URL https://api.semanticscholar.org/CorpusID:215737187.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franccois Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:220250819.

O. Khattab and Matei A. Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. URL https://api.semanticscholar.org/CorpusID:216553223.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2017. URL https://api.semanticscholar.org/CorpusID:2593903.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. URL https://api.semanticscholar.org/CorpusID:86611921.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Jimmy J. Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Nogueira, and David R. Cheriton. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. URL https://api.semanticscholar.org/CorpusID:235366815.

Songshuo Lu, Hua Wang, Yu Rong, Zhi Chen, and Yaohua Tang. Turborag: Accelerating retrieval-augmented generation with precomputed kv caches for chunked text. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:273233795.

Manas Mehta, Fangcong Yin, and Greg Durrett. Randomized yarn improves length generalization for long-context reasoning. *ArXiv*, abs/2606.23687, 2026. URL https://api.semanticscholar.org/CorpusID:289629978.

William Merrill, Yanhong Li, Tyler Romero, Anej Svete, Caia Costello, Pradeep Dasigi, Dirk Groeneveld, David Heineman, Bailey Kuehl, Nathan Lambert, Jacob Morrison, Luca Soldaini, Finbarr Timbers, Pete Walsh, Noah A. Smith, Hanna Hajishirzi, and Ashish Sabharwal. Olmo hybrid: From theory to practice and back. *ArXiv*, abs/2604.03444, 2026. URL https://api.semanticscholar.org/CorpusID:287199770.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *ArXiv*, abs/2309.00071, 2023. URL https://api.semanticscholar.org/CorpusID:261493986.

Jason Priem, Heather A. Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*, abs/2205.01833, 2022. URL https://api.semanticscholar.org/CorpusID:248512771.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27:129–146, 1976. URL https://api.semanticscholar.org/CorpusID:45186038.

Shreya Shankar, Aditya G. Parameswaran, and Eugene Wu. Docetl: Agentic query rewriting and evaluation for complex document processing. *ArXiv*, abs/2410.12189, 2024. URL https://api.semanticscholar.org/CorpusID:273374845.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han Wang, Hai-Bo Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Ö. Arik, Danqi Chen, and Tao Yu. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *ArXiv*, abs/2407.12883, 2024. URL https://api.semanticscholar.org/CorpusID:271270735.

Diane Tchuindjo, Devavrat Shah, and Omar Khattab. Obliq-bench: Exposing overlooked bottlenecks in modern retrievers with latent and implicit queries. *ArXiv*, abs/2605.06235, 2026. URL https://api.semanticscholar.org/CorpusID:288014448.

Llama Team. The llama 3 herd of models, 2024.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663, 2021. URL https://api.semanticscholar.org/CorpusID:233296016.

Barbara B. Tillett. A taxonomy of bibliographic relationships. *Library Resources & Technical Services*, 35:150–158, 1991. URL https://api.semanticscholar.org/CorpusID:59748210.

John W. Tukey. The future of data analysis. *Annals of Mathematical Statistics*, 33:1–67, 1962. URL https://api.semanticscholar.org/CorpusID:122864799.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeff Han, Isa Fulford, Hyung Won Chung, Alexandre Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *ArXiv*, abs/2504.12516, 2025. URL https://api.semanticscholar.org/CorpusID:277857238.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *ArXiv*, abs/2508.21038, 2025. URL https://api.semanticscholar.org/CorpusID:280949957.

Emily Xiao, Chin-Jou Li, Yilin Zhang, Graham Neubig, and Amanda Bertsch. Efficient many-shot in-context learning with dynamic block-sparse attention. *ArXiv*, abs/2503.08640, 2025. URL https://api.semanticscholar.org/CorpusID:276928367.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *ArXiv*, abs/2412.06464, 2024. URL https://api.semanticscholar.org/CorpusID:274598177.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://api.semanticscholar.org/CorpusID:52822214.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *ArXiv*, abs/2410.02694, 2024. URL https://api.semanticscholar.org/CorpusID:273098808.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062, 2020. URL https://api.semanticscholar.org/CorpusID:220831004.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *ArXiv*, abs/2512.24601, 2025. URL https://api.semanticscholar.org/CorpusID:284350669.

A Limitations

While this work shows initial evidence for challenging properties of CTC tasks, there is still room for future work to expand on our analysis. Our paper, while supporting our hypotheses up to 32k tokens in-domain (128k tokens in our length generalization experiments) due to compute limitations does not examine larger corpus-scale settings beyond 128k tokens—note that for many high-CTC tasks performance already degrades greatly by 32k tokens and shows clear trends of degrading further. While we aim to get a comprehensive set (including 12 tasks representative of previously explored settings), there are still likely several low-CTC and high-CTC tasks our suite misses, and that we encourage future work to examine and reproduce our findings on. While this work helps define and understand high-CTC tasks, it only take initial steps towards solving them, and we are excited to see future work make further progress on high-CTC tasks.

B Discussion of Real High CTC Tasks

A key message of this work is to encourage more work on $O_T(N^2)$ and $O_T(NM)$ tasks, which represent very different architectural challenges at scale. While our investigation focuses largely on the technical and architectural aspects of this problem, we emphasize that while these tasks are under-explored, they *are* realistic, and systems that can do such tasks over large corpora may open up various new valuable applications not previously possible, inspiring our current task suite. We below detail just a few applications and potential ideas for future work to investigate that could benefit from such systems at scale, connected to tasks from our suite:

Contradiction / Redundancy Detection (high-CTC tasks a) Motivating example: scientific literature. Given a large and growing body of scientific literature, a key and evolving question in determining new research and experiments is always *what information does the community have?* Often, one will examine a specific research area at a local level, and given specific ideas may, with keyword search and other tools, try to determine what's been explored and not, or what's been verified or not, which is $O_T(N)$. To determine the state of an entire field at a global level (e.g. all claims with the most contradictory evidence, claims with the most redundant evidence), allowing for systematic synthesis and resolution of such evidence, is $O_T(N^2)$, and not currently tractable. Better corpus reasoning systems could allow such systematic analysis end-to-end for scientific literature, fact-checking, minimizing / tracking redundancy in long-text like books, and many other applications.

Discovering long-tail categories / phenomena (high-CTC tasks f, g) Motivating example: Monitoring LLM Generations. Language models can generate text at rapid scale, and for an LLM provider or other larger-scale user, every day a new corpus may get generated. For such instances, being able to faithfully look for the long tail of the least common events with harmful or other properties may be important to the safe deployment of such systems (e.g. *What are the weirdest questions today?*). Doing so however is $O_T(NM)$ where M is the number of possible phenomena (which in the long-tail case will be quite large), and so doing such a task at high-fidelity may currently be very challenging. Similar analyses could be applied for finding under-explored research areas / methodologies.

Structured Re-Organization of Large Corpora (high-CTC task i) Corpora are often unstructured and, aside from some metadata, relatively unorganized. Being able to do analyses like sorting passages in a book by importance, or hierarchically re-organizing and grouping a corpus into a fashion reflecting the overall semantic structure (e.g. to reflect new trends in the corpus) can all be $O_T(NM)$ or $O_T(N^2)$, and potentially provide valuable analyses or new ways of interacting with corpora.

Cross-Corpus Comparison (high-CTC tasks b, c, d, e) Often corpora don't exist as monoliths in isolation. We may be interested in understanding how corpora evolve over time from largely similar distributions with high overlap (e.g. historical records, internet text, internal LLM agent traces from different live model checkpoints), or we may be interested in finding connections between unrelated corpora (for example two unrelated fields of science). These tasks are versatile, and in large part often end up being $O_T(N^2)$ as a function of the corpora being compared.

We believe that if systems which can handle such tasks become more scalable, given that the above are often difficult (or intractable) and may rely on a lot of manual work and heuristics with current systems, a large variety of applications may emerge from new methods here. Work towards building realistic benchmarks for these, as well as new architectures that can handle them, are valuable future directions.

C Task Suite Data Details

C.1 Data Construction

Evaluation Set Construction We evaluate at lengths up to 128k tokens (our code supports up to 10M tokens for several tasks). For each task we typically fix a canonical query set (usually 500 examples per set), and then will across different length evaluation sets within the task expand the distractor set. The documents and queries between train and eval sets are by default kept completely disjoint to avoid any sort of contamination.

Dataset	Example
O(N): One-pass tasks	
NQ	*"who sold out jesus for 30 pieces of silver"* → `[1]`
HotpotQA (bridge)	*"What company did Rex Maughan aquire?"* → `[8, 9]`
NIAH-contra	*"... a hypertonic solution of 14.4% has an **ir**reversible ciliostatic effect"* → `[20]`, the one claim saying *reversible*
BEIR SciFact	*"0-dimensional biomaterials show inductive properties."* → `[1]`
BEIR FiQA	*"Where should I park my rainy-day / emergency fund?"* → `[2, 8, 10, 12, 19]`
MS MARCO	*"what is the difference between fables and folktales? for kids"* → `[2]`
MS MARCO rerank	the same 20-passage pool, every passage scored → `Ranking: [2], [1], [3], ...`
OBLIQ	*"... written in a similar style and voice to: The slowness and calmness of the lodge came with genuine inquisitiveness..."* → `[14, 23, 29]`
OOLONG	10 labelled sentence pairs; *"which of the labels is the most common?"* → `Label: incorrect`
Outlier (Amazon)	20 reviews, 16 rated 4-star and 4 rated 5-star → `Outliers: 4, 7, 11, 17`
Outlier (wiki, fixed M)	57 chunks from 3 articles (43 / 11 / **3**) → `Outliers: 5, 17, 54`
Absence	90 numbered sentences, 3 deleted from the copy → `["The interior is what", "In opposite chapels are", ...]`
Higher-complexity tasks	
Outlier (wiki, scale-k)	55 chunks from 10 articles (9 / 7 / 7 / ... / **3**) → `Outliers: 26, 41, 42`
Grouping (OpenAlex)	*"Cluster these 20 papers into 15 groups."* → `{"groups": [{"doc_ids": [4, 5, 12]}, ...]}`
Contradiction	`[2]` *"... has a reversible ciliostatic effect"* vs. `[12]` *"... has an irreversible ciliostatic effect"* → `[[2,12], [8,11], [13,17]]`
xabsence	19 claims, 8 matched pairs, 3 with no twin in the other corpus → `Unmatched: 1, 3, 14`
strmatch	`[1]` and `[11]` share *"kossuth westward taco"* → `[[1,11], [2,8], [15,19]]`
qdmatch (NQ)	`[7]` *"when did 10 shilling note go out of circulation"* ↔ `[29]` → `[[7,29], [14,28], [19,40]]`
qdmatch (HPQA)	`[8]` *" 'Lost!' is a song by a British rock band formed in what year?"* ↔ `[25]` and `[34]`
qdmatch (FiQA)	`[25]` *"Pay off credit card debt or earn employer 401(k) match?"* ↔ `[55]`
Reorder	50 shuffled segments of *Flatland* → `[31, 19, 12, 22, 37, ...]`
Textgroups	noun counts $34{+}31{+}5 = 70$ and $13{+}17{+}40 = 70$ → `[[1,11,12], [2,14,17]]`

Table 2: **One real example per task.** All content is verbatim from the evaluation sets, elided with ... where a passage is too long to show; document IDs are 1-indexed exactly as the model sees them. Appendix H shows each example in full, with the task instruction, the surrounding corpus and the distractor types.

Training Data Construction Unless stated otherwise, each task is trained for one epoch on 20,000 examples split evenly across five context budgets (so 4,000 examples for 2k, 4k, 8k, 16k and 32k respectively, so the training distribution matches the evaluation uniformly. Training examples are drawn from the same generator as the evaluation examples for that task, and for the tasks with a published split (e.g. NQ, etc.), we use existing train/test divisions. OBLIQ is the sole exception: it has no usable training split and its training data is synthesized, though for evaluation we use its existing evaluation set with BM25 mined hard negatives.

C.2 Tasks: Overview

C.2.1 $O_T(N)$ tasks: one pass over the corpus

Factoid QA (NQ). *Gold-ID F1.* We use NaturalQuestions-Open (sourcing train and eval from existing train / eval set) [Izacard et al., 2022], with all documents — gold, hard negative and filler — drawn from the 100-word Wikipedia DPR corpus [Karpukhin et al., 2020] served by pyserini [Lin et al., 2021], so every document shares one surface format. For each question we BM25-search [Robertson and Jones, 1976] for a passage containing an answer string (that becomes the gold), take top BM25 hits that do *not* contain the answer as hard negatives, and fill the remaining slots with random wiki-DPR passages. By default we hold hard negatives to ∼10% of the pool.

Multi-hop QA (HotpotQA). *Gold-ID F1.* The bridge subset of HotpotQA [Yang et al., 2018], where answering requires composing two gold passages; the corpus is again wikipedia-100w with BM25 hard negatives. Because the answer needs a fixed (two) number of retrieval steps, this remains $O_T(N)$. We source train examples from the standard train set and test examples from the validation set. We use `cross-encoder/ms-marco-MiniLM-L-6-v2` with a CE filter of 3 as well to remove false positives. And similar to NQ we hold hard negatives to ∼10% of the pool.

BEIR SciFact / FiQA. *Set-F1.* We use two BEIR Thakur et al. [2021] tasks: SciFact (scientific claim → abstract, single-gold) and FiQA (financial opinion QA, median 2 golds/query). Both mine distractors by BM25 from the dataset's own corpus. FiQA is sparsely judged, so plain BM25 distractors to filter unlabeled positives we score (query, candidate) pairs with a cross-encoder and keep a candidate only if it scores at least a margin below the gold, then fill the rest of the pool with random real corpus documents. FiQA's train-set is too small (5,500 examples) so we re-use queries across context lengths to get to 20,000. SciFact's existing train set was too small (809) to go to 20,000 total, so it has a smaller train set (4045 examples, each query used once per length). For SciFact the eval set is only 299 examples so our set doesn't go up to 500. We use the same CE filter pipeline as HotpotQA to remove false positives.

MS MARCO retrieval and reranking. *Set-F1; MRR@10 for docs with positive CE score* Passages come from the `msmarco-v1-passage` index. Hard negatives are taken from SentenceTransformers' precomputed BM25+dense mining and filtered by a margin on precomputed cross-encoder scores (same as above), so we don't re-run cross-encoder here; random fill is sampled as random pids from the 8.8M collection, which are byte-identical in format to gold and hard passages. The rerank variant reuses the identical pool but asks for the documents in ranked order based on cross-encoder score.

Needle contradiction (NIAH-contra). *Set-F1.* An $O_T(N)$ control derived from the $O_T(N^2)$ contradiction task: one member of a gold contradiction pair is promoted to the query ("*find the document that contradicts this claim*") and the other stays in the corpus as the needle. Every other claim, including the other gold pairs, is a distractor. Comparing this against contradiction isolates the effect of all-pairs search from the effect of PubMed claim text.

OBLIQ retrieval. *Set-F1.* OBLIQ-Bench (`dianetc/OBLIQ-Bench`) poses subjective, long-form queries over its own corpora; we port it in-context by forcing all qrel positives into the example and filling with BM25-mined distractors from the same subset. Specifically, we use the twitter subset of the benchmark ("*find tweets where users are implicitly insinuating that great powers quietly profit from sustained turmoil…*"), satisfied by a *set* of documents: median 5 golds per example, mean 9. The documents themselves are single tweets — a median of 257 characters.

Synthetic training data. As OBLIQ has no usable training split: the benchmark has only a few hundred real queries per subset. We thus generate data as follows. For each seed document D (taken from corpus):

1. An LLM (Qwen3-14B-Instruct) writes an oblique query Q that D satisfies, few-shot primed with real (document, query) pairs (held out from final eval set).
2. An *obliqueness* filter rejects any Q with word-Jaccard overlap against D above a per-subset cap (0.55 for the writing subset, 0.35 elsewhere), so a keyword shortcut cannot separate the golds — matching real OBLIQ, whose queries are deliberately non-lexical.
3. BM25 retrieves candidate documents for Q from the same subset.
4. A *multi-gold judge* (also using Qwen3-8B) asks, for each top candidate, whether it *also* satisfies Q. These either become additional golds or hard negatives.

Only training data is synthetic — the held-out evaluation set is the twitter subset of the benchmark (we use all subsets for generation of the training data). Our data is very initial and it's likely that much better data could be generated for this task.

Oolong. *Partial-credit score.* A port of OOLONG [Bertsch et al., 2025]: a long context of labeled classification items, each tagged with a date and a user id, and a distributional question — most/least common label, most frequent user, when label A overtook label B. We use the *unlabeled* context variant, so the model must classify each item and then aggregate. Because the label set is fixed and each item is judged independently before a single aggregation step, this is $O_T(N)$.

Absence. *Set-F1 over removed IDs.* Following AbsenceBench [Fu et al., 2025]: the model sees a full numbered corpus, then a second copy with some elements deleted, and must name what is missing. We construct data with random Project Gutenberg passages, with sentence-level chunks: each document is deleted independently with probability p. Because the second copy preserves the original order, the two corpora are positionally aligned.

Outlier detection (Amazon Reviews). *Set-F1.* Each corpus contains customer reviews from the Amazon Reviews 2023 corpus [Hou et al., 2024], shown as review title and body only. Examples are mixed 50/50 between two variants. In the *rating* variant a majority of reviews share one star rating and 3 outliers carry a different one (*"find the outliers with the least common rating"*); in the *category* variant the majority share one product category and 3 outliers come from another (*"find the outliers from the least common category"*), with categories taken from the corpus's own product-category labels.

C.2.2 $O_T(NM)$ tasks: categorization

Outlier detection (Wikipedia). *Set-F1.* Each context is filled with 100-word chunks from a handful of Wikipedia articles with imbalanced counts; the answer is the ID list of chunks belonging to the article with the fewest chunks. We run two variants: *scale-M*, where the number of source articles grows with N ($M = N/5$ on average), and a *fixed-M* control where the article count stays constant (at a value of 3) as N grow.

Grouping (OpenAlex). *Kendall Tau with true order labels* Given N scientific abstracts, partition them into M groups. Gold partitions come from the OpenAlex [Priem et al., 2022] concept hierarchy: we sample only papers annotated down to level L3 (lower levels indicate more specific categories), then per example pick a level $L \in \{L0, \ldots, L3\}$ and sample M distinct concept values at that level. The level sets the granularity—L0 gives coarse groups to L3 fine ones—and the model must infer from the corpus and M what granularity is being asked for. Note that M will naturally be larger for larger corpora with these heuristics.

C.2.3 $O_T(N^2)$ tasks: all-pairs search

Cross-corpus absence (xabsence). *Set-F1.* Two corpora A and B share near identical document sets. Almost every document in one has a twin in the other; exactly k are unmatched, and the model names these. The suite uses an **exact-copy** variant: a twin is an identical copy of its partner, and documents are full PubMed abstracts rather than single sentences, but this task could be made more challenging with paraphrasing or other variants.

Query-document matching (qdmatch). *Pair set-F1.* The retrieval analogue of contradiction for 3 retrieval tasks. Pools are drawn so that relevant questions, distractor questions (gold withheld) and distractor documents (questions withheld) are disjoint such that only three pairs of (question, document) are in the corpus. The model must find the sparse gold pairs among $N \times N$ combinations. We build three versions from three retrieval sources—NQ (1 gold/question), HotpotQA bridge (2 golds/question), and FiQA (median 2 golds/question)—so difficulty from the underlying retrieval problem varies while the search structure is fixed. Note that there are no hard negatives here, but the process to get train data and sources are the same as the original retrieval sets described above.

Contradiction search (PubMed). *Pair set-F1* Each document is a single claim sentence from a PubMed abstract [Jin et al., 2019]. Among N documents, $K{=}3$ gold pairs (d, d') are hidden: d is a real PubMed sentence and d' is a sentence, written by Qwen3-14B-Instruct, reporting what a *different* study might have found such that the two cannot both be true (and prompted with the style of other claim pairs). A word-overlap threshold (Jaccard ≤ 0.5) drops any near-duplicates. Only one sentence per pair is LLM-written. Candidate sentences are filtered via regex to be self-contained and claim sentences, and fillers come from abstracts disjoint from every gold-source abstract in that example. While we manually inspect 100 examples for obvious artifacts, subtle LLM-generated annotation artifacts likely make this version of the task much easier than more realistic versions of contradiction search, which we encourage future work to investigate.

String matching (strmatch). *Pair set-F1.* N strings of L words drawn from a Wikipedia vocabulary, and the task is to find every pair sharing a contiguous run of $\geq k$ words. Gold pairs share exactly one k-word run; hard negatives share a $(k{-}1)$-word run. This task has no semantic content at all, so the only thing that scales is the number of comparisons.

Reordering. *Kendall-τ.* N consecutive sentence segments of a single Project Gutenberg book, presented in random order, with the task of recovering the original ordering as a list of document IDs. Because sorting requires comparisons between arbitrary pairs of segments, we label it $O_T(N^2)$ (unlike number sorting we can't assume a strict global positional prior).

C.2.4 Beyond $O_T(N^2)$

Textgroups. *Group set-F1.* N chunks (sourced from Project Gutenberg), each carrying a feature (how many nouns / verbs / adjectives it contains, or how often a chosen connective appears), and the task is to find every triple whose feature values sum to a target T. Because $G{=}3$, the construction can both plant exactly K triples and verify by brute force that no others exist. This is the one task in the suite where the per-document oracle operation is itself nontrivial (counting) *and* the CTC is greater than quadratic. We include it as an example of tasks that can go beyond $O_T(N^2)$.

D Data Statistics / Evaluation

Table 3 lists, for every task in the suite, the corpus it is built from, its CTC class, the number of documents present at each context rung, and the number of examples it is evaluated on. Each example carries `documents` (a list of {title, text}), `queries`, `answers`, and `gold_doc_indices`. Document counts are measured from the evaluation files. Document length is given in characters.

Evaluation We typically serve everything with vLLM Kwon et al. [2023] (we use a modified version for block-sparse inference), where block-sparse attention is implemented by masking with respect to marker tokens. For generation we use greedy decoding.

				Documents per example (N)						
Task	**Source corpus**	**CTC**	**Metric**	2k	4k	8k	16k	32k	**Chars/doc**	**Eval**
NQ	Wikipedia 100w (DPR)	$O_T(N)$	gold-ID F1	11	23	48	104	208	636	500
HotpotQA (bridge)	Wikipedia 100w (DPR)	$O_T(N)$	gold-ID F1	17	36	72	144	296	432	500
NIAH-contra	PubMed claims	$O_T(N)$	gold-ID F1	40	86	180	365	740	60	500
BEIR SciFact	SciFact abstracts	$O_T(N)$	gold-ID F1	5	10	21	43	88	1517	300
BEIR FiQA	FiQA-2018 posts	$O_T(N)$	gold-ID F1	8	19	40	82	166	835	500
MS MARCO	msmarco-v1-passage	$O_T(N)$	gold-ID F1	23	46	93	187	374	337	500
MS MARCO rerank‡	msmarco-v1-passage	$O_T(N)$	MRR@10	20	40	70	100	376	336	500
OBLIQ	OBLIQ-Bench posts	$O_T(N)$	gold-ID F1	3	6	30	19	30	4196	126
OOLONG§	labeled item stream	$O_T(N)$	partial credit	41	83	153	337	616	154	500
Outlier (Amazon)	Amazon Reviews 2023	$O_T(N)$	set-F1	20	40	80	160	320	494	500
Outlier (wiki, fix-M)	Wikipedia 100w	$O_T(N)$	set-F1	14	28	57	111	220	612	500
Absence	Project Gutenberg	$O_T(N)$	set-F1	90	180	360	720	–	146	500
Outlier (wiki, scale-k)	Wikipedia 100w	$O_T(NM)$	set-F1	14	28	57	111	220	612	500
Grouping	OpenAlex abstracts	$O_T(NM)$	pairwise-F1	10	21	43	88	176	893	500
Textgroups	synthetic passages	$O_T(N^3)$	group-F1	11	24	50	103	210	1182	500
Contradiction	PubMed claims	$O_T(N^2)$	set-F1	56	92	187	379	762	148	500
Cross-corpus absence	PubMed claims	$O_T(N^2)$	set-F1	39	81	165	333	669	139	500
String matching	synthetic strings	$O_T(N^2)$	set-F1	38	82	170	350	700	81	500
QDmatch (NQ)	Wikipedia 100w (DPR)	$O_T(N^2)$	pair-F1	18	40	84	174	356	338	500
QDmatch (HotpotQA)	Wikipedia 100w (DPR)	$O_T(N^2)$	pair-F1	18	40	84	174	356	272	500
QDmatch (FiQA)	FiQA-2018 posts	$O_T(N^2)$	pair-F1	10	26	58	122	248	548	500
Reordering	Project Gutenberg	$O_T(N^2)$	Kendall tau	12	27	57	116	–	581	500

Table 3: **Corpus statistics for every task in the evaluation suite** Column headings 2k–32k are approximate token targets per task; the entries are the median number of documents actually present in an example of that rung, measured from the evaluation files the reported results were graded on. Chars/doc is the median document length in characters at the deepest available rung.

E Additional Experiments

E.1 Model Scale

We include a more detailed plot with some different dense and block-sparse attention values for our model scale comparison (Figure 7). This includes dense attention model scale results on Reordering and QDMatch (NQ), where we find a trend of smaller model scales sometimes being unable to learn challenging high CTC tasks.

E.2 Model Family

We in Figure 8 show a few exact model family performance numbers with both block-sparse and full attention. Note that block-sparse struggles greatly on contradiction across model families.

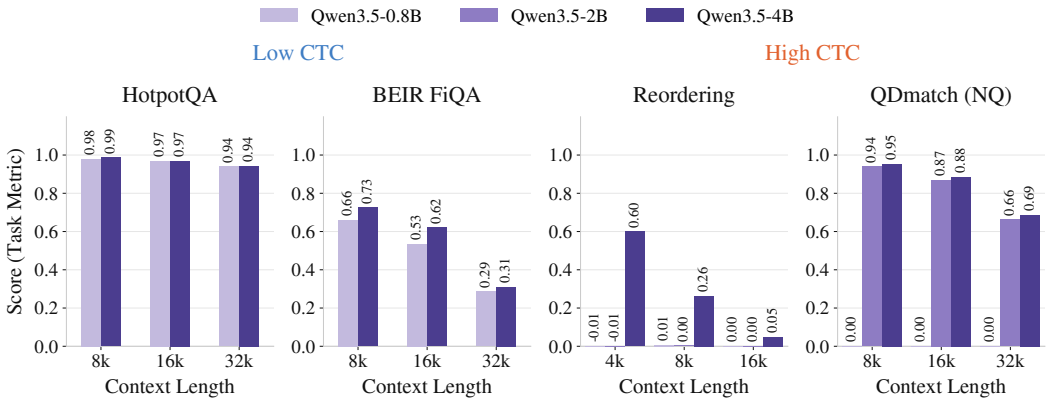

Figure 7: Full model performance on several tasks and Qwen3.5 model scales. Some high-CTC tasks have minimum model scales necessary to learn effectively.

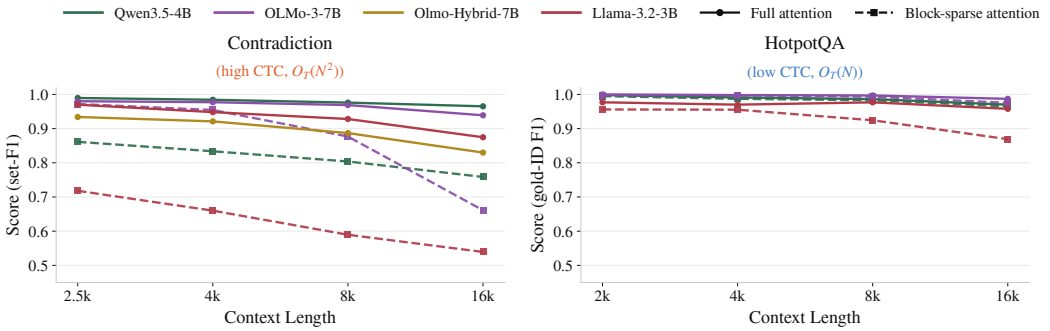

Figure 8: Full vs block-sparse attention with a few different model families.

F Training Hyperparameters

All CTC Suite results are full-parameter fine-tunes of a base checkpoint, trained with `olmo-core`. Every arm of a comparison — full attention, block-sparse, and block-sparse with mask mixing — is trained from the same base with the same recipe, so the only difference between arms is the attention mask used during training. Defaults shared across all settings:

Optimizer: AdamW with skip-step (a step whose loss or gradient norm is a large outlier is skipped rather than applied), lr 5×10^{-5}, $\beta = (0.9, 0.95)$, weight decay 0, gradient clipping at global norm 1.0. **Schedule:** linear warmup over the first 3% of steps, then linear decay to 0. **Epochs / batch:** 1 epoch, global batch 8 sequences, 1 instance per micro-batch. **Precision:** bf16 parameters with fp32 gradient reduction, FSDP for data parallelism, activation checkpointing on. **Packing:** examples are *not* packed; one example per sequence, so per-example sequence lengths are stable across runs and document boundaries stay aligned with chunk boundaries.

Mask mixing. We apply mask-mixing as a *curriculum* from 0.8 to 0 rather than a fixed p, finding this to work well empirically. Evaluation always uses the block-sparse mask with no mixing, so a mask-mixing result reflects pure block-sparse inference and differs from the pure-block-sparse arm only in how it was trained.

G GPU Resources / Compute

The experiments for our investigation involved roughly 3k A100 hours and 1k H200 hours (32 CPU cores, 300GB RAM). We expect that with 8xH100 GPUs, the results in the paper Figure 3 would take under 500 hours to reproduce.

H Task Examples

We show one real example for every task in the suite (Table 1), we omit full corpus text omitted for readability.

H.1 $O_T(N)$ tasks

H.1.1 NQ (factoid retrieval)

Instruction: *"Use the given documents to identify which document is most relevant to answering the question. Write your answer in the following format: Relevant Document: [id]"*
Query: *"who sold out jesus for 30 pieces of silver"*
Context (3 of 20 documents shown):

> `[1]` *(gold)* "Thirty pieces of silver was the price for which Judas Iscariot betrayed Jesus, according to an account in the Gospel of Matthew 26:15 . . . "
> `[2]` *(BM25 hard negative)* "Butuan Silver Paleograph. The Butuan Silver Palaeograph, also known as the 'Butuan Silver Strip', is a piece of metal with inscriptions . . . "
> `[3]` *(BM25 hard negative)* "Phantom Stranger . . . of betraying him. As the Spectre is about to attack the Stranger, a Mysterious Voice sends him off . . . As payment for what occurred . . . "
> `[· 17 more documents omitted]`

Gold doc ID: `1` **Answer string (not the target):** *"Judas Iscariot"*

H.1.2 HotpotQA (bridge, multi-hop retrieval)

Instruction: *". . . identify which documents are relevant . . . Relevant Documents: [id1], [id2]"*
Query: *"What company did Rex Maughan aquire?" (question text verbatim, including the source typo)*
Context (4 of 20 documents shown):

> `[8]` *(gold-1)* "Forever Living Products International, Inc. (FLPI) is an American privately-held multi-level marketing (MLM) company based in Scottsdale, Arizona . . . "
> `[9]` *(gold-2)* "Rex G. Maughan is an American businessman. He is the founder, president, and chief executive officer of Forever Living Products . . . "
> `[1]` *(hard negative — name collision)* "Alex Maughan (born April 24, 1995) is an American rugby union player who plays in the front row for the United States men's national team . . . "
> `[2]` *(hard negative — same article, wrong chunk)* "Ruth graduated from BYU with a degree in elementary education, she and Rex moved to Arizona . . . "
> `[· 16 more documents omitted]`

Gold doc IDs: `[8, 9]` **Answer string:** *"Aloe Vera of America"*

H.1.3 NIAH-contradiction ($O_T(N)$ control for contradiction)

Instruction: *single-document retrieval, as in NQ.*
Query: *"Find the document that directly contradicts the following claim: The National Lung Screening Trial, performed mainly in academic medical centers, showed that cancer mortality can be reduced with computed tomography (CT) screening compared with chest radiography in high-risk patients."*
Context (4 of 40 documents shown; the 2k rung):

> `[23]` *(needle)* "A large-scale randomized trial indicated that, for individuals at elevated risk, chest X-rays were more effective than CT scans in lowering cancer-related deaths."
> `[1]` *(filler)* "Thirteen focus group discussions involving a total of 97 informants were conducted."
> `[2]` *(filler)* "TRIM47 expression levels were found to be significantly increased in PC compared to benign tissues by both immunohistochemistry and qRT-PCR . . . "
> `[3]` *(filler)* "Free cortisol was significantly elevated at 8:00 to 8:30 hours in the high job strain group but not at later times of the day or evening."
> `[· 36 more documents omitted, drawn from PubMed abstracts disjoint from the needle's source abstract]`

Gold doc ID: `23`.

H.1.4 BEIR SciFact

Instruction: *single/multi-document retrieval.*
Query (a scientific claim): *"0-dimensional biomaterials show inductive properties."*
Context (3 of 22 abstracts shown):

> `[1]` ***(gold)*** "New opportunities: the use of nanotechnologies to manipulate and track stem cells. Nanotechnologies are emerging platforms that could be useful in measuring, understanding, and manipulating . . . "
> `[2]` ***(BM25 negative)*** "The spectrum of retinopathy in adults with Plasmodium falciparum malaria . . . "
> `[3]` ***(BM25 negative)*** "Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome . . . "
> `[· 19 more documents omitted]`

Gold doc ID: `1`

H.1.5 BEIR FiQA (cross-encoder–cleaned negatives)

Instruction: *multi-document retrieval.*
Query: *"Where should I park my rainy-day / emergency fund?"*
Context (3 of 20 documents shown; 5 golds in this example):

> `[2]` ***(gold)*** "I would suggest your local credit union or local bank for security and liquidity. Liquidity is probably the most important issue for a emergency fund."
> `[8]` ***(gold)*** "This is probably a good time to note that credit is not a liquid asset, and not an emergency fund. Credit can be revoked or denied at any time . . . "
> `[10]` ***(gold)*** "First off, you generally want to park your emergency fund somewhere that is 'safe', meaning something that is not subject to market fluctuations . . . "
> `[· 17 more documents omitted: cross-encoder-surviving hard negatives (∼10% of slots) plus random real corpus documents]`

Gold doc IDs: `[2, 8, 10, 12, 19]`. FiQA is sparsely judged, so candidates scoring close to the gold under the cross-encoder are dropped rather than used as negatives — they are likely unlabeled positives.

H.1.6 MS MARCO (retrieval)

Instruction: *single-document retrieval.*
Query: *"what is the difference between fables and folktales? for kids"*
Context (3 of 20 passages shown):

> `[2]` ***(gold)*** "Folktale vs Fable. Folktales and fables can be understood as two different types of stories that show a difference between them . . . "
> `[1]` ***(hard negative)*** "Your job is to make sense of the fact or example in the context of the overall main idea that is being conveyed . . . "
> `[3]` ***(random fill)*** "Cascamite is a waterproof glue and is probably the must effective glue of all . . . "
> `[· 17 more documents omitted]`

Gold doc ID: `2`. Hard and random passages are drawn from the same index, so they are format-identical: there is no stylistic cue distinguishing a mined negative from a random one.

H.1.7 MS MARCO reranking

Instruction: *"Rank the documents by how relevant each is to the question, from most to least relevant. Output the document IDs in ranked order. Write your answer in the following format: Ranking: [id1], [id2], [id3], . . . "*
Query: *"what is the difference between fables and folktales? for kids"*
Context (5 of 23 passages shown; the 2k rung. Cross-encoder scores not in the actual prompt):

[3] *(CE* $+8.04$*; also the MS MARCO qrel positive)* "Folktale vs Fable. Folktales and fables can be understood as two different types of stories that show a difference between them. Mostly, folktales and fables are passed on from one generation to another orally . . . "
[17] *(CE* $+4.60$*)* "A fable differs from a parable in that the latter excludes animals, plants, inanimate objects, and forces of nature as actors that assume speech . . . "
[12] *(CE* $+4.40$*)* "A fable is a moral tale that often features animal characters. We often associate fables with the master of them all, Aesop . . . "
[9] *(CE* -9.79*)* "A mood disorder is a mental health class that health professionals use to broadly describe all types of depression and bipolar disorders . . . "
[19] *(CE* -11.30*, the lowest in the pool)* "Geology 1003 with Weaver at University of Oklahoma . . . "
[· 18 more passages omitted]

Target (top-10): `Ranking: [3], [17], [1], [12], [9], [22], [13], [16], [14], [5]`

H.1.8 OBLIQ (subjective retrieval, twitter subset)

Instruction: *multi-document retrieval.*
Query: *"Find tweets where users are implicitly insinuating that great powers quietly profit from sustained turmoil in West Asia, suggesting that disruptions at key maritime chokepoints and damage to energy facilities maintain elevated fuel costs and tilt market share toward them, while questioning which actors truly profit from an extended standoff."*
Context (3 of 33 documents shown; 2k rung):

[19] *(gold)* "Yes... So what happens if all of that Middle Eastern infrastructure gets destroyed over the next year or so including Iran's? Who benefits in terms of oil sales? Mostly the US and Russia."
[27] *(gold)* "Who stands to gain and who stands to lose if the Strait of Hormuz stays closed?"
[3] *(distractor)* "Strong conviction on the spike! Brent's holding near $103 spot right now with Hormuz disruptions from the Iran strikes. Jumping to $119 today or $140 by Monday would demand major sustained..."
[· 30 more documents omitted, BM25-mined from the twitter subset's 500-document pool]

Gold doc IDs: `[4, 19, 27, 30, 32]`

H.1.9 OOLONG (aggregation over labeled items)

Instruction: *"Read the data below and answer the question. Compute the exact answer by analyzing every item; do not guess or approximate."*
Context (rendered verbatim as one stream, not as numbered documents; 2k rung, 59 items, one per line):

"The following lines contain 59 general-knowledge questions, one per line. Each question has an answer that can be described as one of 6 categories: 'abbreviation', 'entity', 'human being', 'numeric value', 'location', 'description and abstract concept'.
You will be asked to answer questions about the aggregate label statistics across all 59 examples in this dataset. Do not try to guess, estimate, or approximate the result. Calculate the exact answer given these datapoints.
`Date: Oct 24, 2022 || User: 34487 || Instance:` *Where is Inoco based ?*
`Date: Jul 14, 2022 || User: 11784 || Instance:` *Where did the term " 86 " come from ?*
`Date: Feb 23, 2023 || User: 57329 || Instance:` *What international sport was originally called hurley ?*
`Date: Feb 09, 2024 || User: 24838 || Instance:` *What is SVHS ?*
`Date: Apr 17, 2025 || User: 60511 || Instance:` *What is HDLC ?*
[· 54 more items omitted]"

Query: *"For the following question, only consider the subset of instances that occur in April of any year. Among instances occuring in April, which of the labels is the most common? Give your final*

answer in the form 'Label: answer' . . . "
Answer: `abbreviation`

H.1.10 Outlier (Amazon Reviews)

Instruction: *"You are given a list of product reviews. Most share a common attribute (star rating or product category); a few are outliers with a different value. First state what the majority attribute is and what the outlier attribute is, then list the 1-indexed document IDs of the outliers . . . Outliers: [id1], [id2], . . . "*
Query: *"Can you find outliers with the least common rating in this data?"*
Context (3 of 20 reviews shown; ratings and categories are withheld from the model — only title and body are rendered):

> `[4]` ***(outlier**, 5-star) "Works great, no errors on memtest"* — "This RAM comes in 2 sticks of 4gb . . . and works perfectly in my old PC. Ran memtest with 0 errors . . . "
> `[11]` ***(outlier**, 5-star) "Great Case at a Fantastic Price"* — "I never realized Amazon had their own line of products, AmazonBasics! . . . The case is the perfect size for my new Canon Powershot Camera . . . "
> `[1]` *(majority, 4-star) "Light"* — "Works well. Can be kind of heavy depending on the size of your camera. Easy to install and use . . . "
> `[17 more reviews omitted]`

Target: *"Most reviews are 4-star ratings and the outliers are 5-star reviews."* followed by `Outliers: [4], [7], [11], [17]`. Ground truth comes from structured metadata only (the star-rating field), never from an LLM label. The attribute set is small and fixed — five star ratings, or a closed category list — which is what keeps this variant $O_T(N)$: there is no category to discover.

H.1.11 Outlier (Wikipedia, fixed M)

The $O_T(N)$ control for the scale-k outlier task: identical format and generator, but the number of source articles M is pinned as N grows instead of scaling with it.
Query: *"Can you find passages that are about a different topic than the rest of these passages?"*
Context (3 of 14 chunks shown at the 2k rung; article titles are withheld):

> `[2]` ***(outlier)*** "1987, and 1989. Several notable players played for Sheppard in the 1980s: John Morris, Tony DeFrancesco, Pat Pacillo, Rich Scheid, Craig Biggio . . . "
> `[14]` ***(outlier)*** "Future major leaguers Rick Cerone and Dan Morogiello played for Sheppard during the 1970s. Seton Hall appeared in the NCAA Tournament three more times . . . "
> `[1]` *(majority)* ". . . account of black immigration to the United States from the Caribbean dates back to 1619, when a small group of voluntary indentured workers arrived in Jamestown . . . "
> `[· 11 more chunks omitted]`

Article distribution (M=3, pinned): `Privy Council of Sweden`: 6, `Trinidadian and Tobagonian Americans`: 5, `Mike Sheppard (baseball)`: **3** (outlier). **Target:** `Outliers: [2], [11], [14]`. Holding M fixed while N grows is what separates $O_T(N)$ from the $O_T(NM)$ scale-k row: with a constant article count, one exemplar per topic suffices and no group discovery is needed.

H.1.12 Absence (Gutenberg text-diff)

Instruction: *"Above are two versions of the same passage. Version B is identical to Version A except that some whole sentences have been removed. Identify every sentence that appears in Version A but is MISSING from Version B. For each missing sentence, write its first four words. Write your answer as a JSON list of strings, in order of occurrence . . . "*
Version A (2 of 90 sentences shown; one Gutenberg travelogue, rendered as flowing prose — this task has *no* document IDs):

> "How lonely it makes one to stand still and feel that of all the mighty throng which divides itself around him, not a being knows or cares for him! What knows he too of the thousands

> who pass him by? . . . **The interior is what one would expect to behold, after viewing the outside.** . . . **In opposite chapels are the tombs of Mary and Elizabeth, and near the former that of Darnley.** . . . ”

Version B: the same passage with those three sentences deleted and nothing else changed.
Target: `["The interior is what", "In opposite chapels are", "There is an innocence,"]` — scored by set-F1 over the four-word prefixes. Scattered single deletions are the hard regime; large contiguous gaps are easy.

H.2 $O_T(NM)$ tasks

H.2.1 Outlier detection (Wikipedia, Scale M)

Instruction: *“. . . First state what the majority attribute is and what the outlier attribute is, then list the 1-indexed document IDs of the outliers. Outliers: [id1], [id2], . . . ”*
Query: *“Can you find passages that are about a different topic than the rest of these passages?”*
Context (4 of 55 chunks shown; article titles are withheld from the model):

> `[26]` ***(outlier)*** “. . . inaugural Interactive Agency of the Year award (2006), it gave it to AKQA, recognising the agency’s ‘global culture, creative hires and technological muscle’ . . . ”
> `[41]` ***(outlier)*** “. . . Year for the second year running at the Revolution Awards and Agency of the Year awards from New Media Age and the Interactive Advertising Bureau . . . ”
> `[42]` ***(outlier)*** “In 2014, AKQA won Queen’s Award for Enterprise: Innovation, and was named Most Innovative Agency at 2014 Digiday Awards . . . ”
> `[1]` *(majority)* “. . . teach and children learn. It will offer a continuing look at how new technology such as wikis, blogs, vlogs, RSS, podcasts, social networking sites . . . ”
> `[· 51 more documents omitted]`

Article distribution (M=10 articles): `José Holebas`: 9, `Andy Carvin`: 7, `Devi Ahilya Vishwavidyalaya`: 7, `SMS Blücher`: 6, `Iraqi insurgency`: 5, `La Baie, Quebec`: 5, `Terra (mythology)`: 5, `Mane people`: 4, `Peter Sloterdijk`: 4, `James Hilton (designer)`: **3** (outlier). **Gold doc IDs:** `[26, 41, 42]`
Note that M grows with N in the *scale-k* variant shown here; the fixed-M control holds the article count constant as N grows.

H.2.2 Grouping (OpenAlex)

Instruction: *“. . . Group them into the requested number of categories based on what they are about. Output a JSON object of the form {“groups”: [{“doc_ids”: [. . .]}, . . .]}. Every document must appear in exactly one group.”*
Query: *“Cluster these 20 papers into 15 groups.”* (concept level L=3, the finest granularity; the query states k but never states the axis)
Context (4 of 20 abstracts shown):

> `[1]` *“Microneedles’ Device: Design, Fabrication, and Applications”* — “The delivery of therapeutical molecules through the skin, particularly to its deeper layers, is impaired due to the stratum corneum layer . . . ”
> `[2]` *“Identifying Resilient Communities in Road Networks: A Path-Based Embedding Approach”* — “Effective resilience analysis of road networks is fundamental to building sustainable and disaster prepared cities . . . ”
> `[3]` *“Role of zinc in health and disease”* — “This review provides a concise overview of the cellular and clinical aspects of the role of zinc . . . ”
> `[4]` *“Unveiling Cutting-Edge Developments in Electrocatalytic Nitrate-to-Ammonia Conversion”* — “The excessive enrichment of nitrate in the environment can be converted into ammonia . . . ”
> `[· 16 more documents omitted]`

Target: the labeled partition, {``groups'': [{``label'': ``Ammonia production'', ``doc_ids'': [4, 5, 12]}, ...]} — only the IDs are scored.
Gold partition (15 clusters over 20 documents): {`[4, 5, 12]`, `[2, 9]`, `[3, 15]`, `[14, 20]`, `[19]`, `[6]`, `[10]`, `[8]`, `[17]`, `[16]`, `[18]`, `[1]`, `[11]`, `[7]`, `[13]`}, whose OpenAlex L3 concepts are *Ammonia production, Autoencoder, Zinc deficiency,*

Macrophage, Elliptic PDE, Moxidectin, Pancreatic cancer, Routing protocol, Stock market, Environmental management system, Formate, Fabrication, Task analysis, Nomological network, Object detection.

H.3 $O_T(N^2)$ tasks

H.3.1 Contradiction (PubMed)

Instruction: *"Given the following corpus of numbered claims, identify all pairs of claims that contradict each other. A pair of claims is contradictory if they cannot both be true at the same time. Output your answer as a JSON list of pairs . . . For example: [[1, 4], [3, 7]]"*
Context (4 of 56 claims shown; the 2k rung, K=3 gold pairs):

> `[9]` *(gold pair with `[37]`)* "A meta-analysis of 15 studies indicates a protective effect of cigarette smoking against the development of clear cell renal cell carcinoma (ccRCC)."
> `[37]` "An association between cigarette smoking and increased risk of clear cell renal cell carcinoma (ccRCC) has been established; however, there are limited data regarding the molecular mechanisms that underlie this association."
> `[31]` *(gold pair with `[56]`)* "Circulating IgG anti-CII converted from positive to negative in 13 patients (10.7%) and from negative to positive in 18 patients (14.8%) among 122 patients with RA . . . monitored sequentially at a mean interval of 12.2 months."
> `[56]` "In a cohort of 122 rheumatoid arthritis patients monitored for IgG anti-CII over a mean interval of 13.4 months, circulating IgG anti-CII levels remained stable in 95.6% of patients, with no significant conversion observed."
> `[· 52 more claims omitted; fillers come from PubMed abstracts disjoint from every gold-source abstract in this example]`

Gold pairs: `[[9, 37], [31, 56], [33, 51]]`.

H.3.2 XAbsence (cross-corpus absence, Gutenberg passages)

Instruction (stated after the corpus; k is given explicitly): *"Below are two corpora of numbered passages, A and B. Every passage in corpus A appears again, word for word, in corpus B — except for exactly 3 passages. Those 3 passages from corpus A are MISSING from corpus B. Find them. Only passages from corpus A can be missing; every passage in corpus B also appears in corpus A. Write your answer in the following format: Missing: [id1], [id2], . . . "*
Context (19 items under one shared index — 11 corpus-A passages, then the 8 corpus-B copies in shuffled order; 6 of 19 shown):

> `[1] A:` "I am, as you doubtless begin to suspect, a fairy."
> `[2] A:` "Looking around, he spied a bird with a long, sharp bill lying on the ground."
> `[3] A:` ***(missing from B)*** "Of course the music, its lilt and the steps that their forefathers had footed to it in the olden time, were as little known to these, the London born, as the tongue and ceremonial of old Peru."
> `[4] A:` ***(missing from B)*** "In this number was the name 'Kukloi' from the Greek word Kuklos (Kuklos), meaning a band or circle."
> `[5] A:` ***(missing from B)*** "It showed (1) that the scholars of the church were being influenced by the new learning; but also (2) that a strict reservation was to be enforced . . . "
> `[· 6 more corpus-A passages omitted]`
> `[18] B:` "I am, as you doubtless begin to suspect, a fairy." *(the word-for-word copy of `[1]`)*
> `[· 7 more corpus-B passages omitted, each an exact copy of an A passage, reordered]`

Target: `Missing: [3], [4], [5]`

H.3.3 Stringmatch

Instruction: *". . . identify all pairs of strings matching the criterion below . . . JSON list of pairs."*
Criterion (in the query): *"Find all pairs of strings that contain a run of at least 3 consecutive*

words in common (the same 3 words, in the same order, appearing contiguously in both strings).”
Context (4 of 20 strings shown; L=10 words each):

> `[1]` *(gold pair with `[11]`)* “diasporas **kossuth westward taco** recipes bakr cease panels neotype niddastausee”
> `[11]` “hermitage wass **kossuth westward taco** bolsheviks iryna epiphany allon garabet”
> `[2]` *(gold pair with `[8]`)* “drown traditionalists eateries visconti broadband **amberdeep chisago kindergartens** taglish portraitist”
> `[8]` “kannada utama likert mollusk budge **amberdeep chisago kindergartens** modifying martingale”
> `[· 16 more strings omitted, including 3 hard-negative pairs sharing a run of exactly 2 words]`

Gold pairs: `[[1, 11], [2, 8], [15, 19]]`. Every word not part of a planted run is globally unique in the example, so these are provably the only qualifying pairs.

H.3.4 QDMatch (NQ source)

Instruction: *“Below is a numbered list of items. Each item is labeled either ‘Query:’ or ‘Document:’. A few query-document pairs are relevant: the document answers the query. Identify every relevant pair . . . [[query_id, document_id], . . .]”*
Context (M=20 queries and N=20 documents under one shared numbering, 5 of 40 items shown; `separate` layout = query block then document block):

> `[7]` *Query* “when did 10 shilling note go out of circulation” *(relevant)*
> `[14]` *Query* “who played the original steve mcgarrett on hawaii five-o” *(relevant)*
> `[19]` *Query* “bosnia and herzegovina croatia macedonia and slovenia all used to be parts of” *(relevant)*
> `[28]` *Document* “Steve McGarrett is a fictional character who is the protagonist of CBS’ ‘Hawaii Five-O’. McGarrett is a former United States Navy officer . . . ”
> `[29]` *Document* “Banknotes of the pound sterling . . . 10 shilling note was designed, featuring Sir Walter Raleigh, which would become the 50 pence note upon decimalisation . . . ”
> `[· 35 more items omitted: 17 distractor queries whose gold documents were withheld, and 17 distractor documents whose queries were withheld]`

Gold pairs: `[[7, 29], [14, 28], [19, 40]]` — 3 relevant pairs among $20 \times 20 = 400$ combinations. Because query pools and document pools are drawn disjointly, the planted pairs are the only true matches.

H.3.5 QDMatch (HotpotQA source)

Same format; each relevant query is a bridge question with *two* gold documents, so k=3 relevant queries yield 6 gold pairs:

> `[8]` *Query* “ ‘Lost!’ is a song by a British rock band formed in what year?”
> `[25]` *Document* “ ‘Lost!’ is a song by the British rock band Coldplay. The band co-produced it with Brian Eno and Markus Dravs for their fourth album . . . ”
> `[3]` *Query* “Which team’s 2013-2014 season had players including a Slovenian who plays at both the point guard and shooting guard positions?”
> `[26]` *Document* “Goran Dragic (born 6 May 1986) is a Slovenian professional basketball[er] for the Miami Heat . . . He plays at both the point guard and shooting guard positions . . . ”
> `[· 36 more items omitted]`

Gold pairs: `[[2, 29], [2, 36], [3, 26], [3, 39], [8, 25], [8, 34]]`

H.3.6 QDMatch (FiQA source)

Same format over FiQA, where queries are financial questions and documents are forum answers — sparsely judged, so candidates scoring close to gold under the cross-encoder are dropped rather than used as negatives:

[2] *Query* "What typically happens to unvested stock during an acquisition?"
[6] *Document* ""I worked for a small private tech company that was aquired by a larger publicly traded tech company. My shares were accelerated by 18 months, as written in the contract. I excer . . . " *(answer text verbatim, including the source typo)*
[· 8 more items omitted]

Gold pairs: `[[2, 6], [3, 9], [4, 8]]`

H.3.7 Reorder (Gutenberg)

Instruction: *"You are given a list of text passages presented in a random order. They were originally consecutive segments of a single document. Output the permutation that restores them to their original order, as a JSON array of the passage IDs."*
Context (3 of 50 segments shown, all from one book — here *Flatland* — in shuffled order; chapter headings are stripped):

[1] "Polygon of two or three hundred sides sometimes — by no means always, for the process is attended with serious risk — but sometimes overleaps two or three hundred generations . . . "
[2] "I had but one voice, and that I had not been aware that his Royal Highness had two. 'That confirms my impression,' said the King, 'that you are not a Man, but a feminine Monstrosity' . . . "
[3] "Linelander. Only by the sound of the voice could sex or age be distinguished . . . "
[· 47 more segments omitted]

Gold ordering: `[31, 19, 12, 22, 37, 21, 32, 14, 33, 48, 41, 46, 4, 25, 36, 20, 42, 18, 15, 23, 1, 40, 6, 7, 35, 16, 50, 28, 27, 39, 9, 30, 43, 3, 34, 2, 24, 47, 49, 45, 13, 29, 44, 10, 26, 5, 38, 11, 8, 17]` (scored by Kendall-τ against the true permutation).

H.4 $O_T(N^3)$ tasks and beyond

H.4.1 Textgroups

Query: *"Each passage has a value: the number of nouns. Find every group of 3 passages whose values add up to 70 (exactly 70)."*
Context (3 of 20 passages shown; the value is a property of the prose, never printed):

[1] *(gold; 34 nouns)* "The castle and the scholar arrived calmly. The river gathered restlessly. The teacher whispered calmly. The saddle whispered calmly. The meadow and the bridge and the garden waited abruptly . . . "
[11] *(gold; 31 nouns)* "The valley explored softly. The beacon watched slowly. The beacon and the harvester and the telescope studied suddenly. The baker guarded wearily . . . "
[2] *(gold of the second triple; 13 nouns)* "The forest and the tower and the compass circled suddenly. The teacher and the bridge faltered bravely. The weaver collapsed wearily . . . "
[· 17 more passages omitted]

Gold groups: `[[1, 11, 12], [2, 14, 17]]` — noun counts $34{+}31{+}5 = 70$ and $13{+}17{+}40 = 70$. The noun / verb / adjective lexicons are closed and pairwise disjoint, so each passage's count is unambiguous, and sentence structure is varied so the count is not a proxy for passage length.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we claim that higher CTC tasks require more computation, and have experimental suggesting approaches with lower computational complexity (Section 4.1)

indeed degrade much faster on tasks with higher complexity per our definition. Our mask mixing approach also, in Table 5, is shown to be quite effective.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, this work does discuss limitations, we acknowledge that there are classes of CTC that we don't cover in the scope of the work. We also acknowledge several times in the work that while our mask mixing approach, while effective, isn't a permanent scalable solution to corpus reasoning, and in fact encourage future work to try to improve upon our initial approaches in this area of research.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper is an empirical paper, our results are experimental.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we include detailed descriptions of the way that our attention masks are implemented, task construction, examples of data in the appendix, and all our data statistics and hyper-parameters also in the Appendix. We also plan to release our data / models in public versions of the paper to further improve reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we include a pointer to an open source repository which should allow reproduction of all of our experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, see Table F, as well as sections on data statistics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: While we don't directly report error bars, we maintain eval sets of at least size 400 (typically 500), and validate consistency of select results across training seeds. All the margins we make claims based on are typically well beyond the margin of uncertainty (e.g. 0.1, 0.2 gaps in metrics).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we include these details in Section G

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, there were no human participants, data concerns, and the work doesn't apply any direct societal harms. In fact, part of the motivation of the work is that larger scale corpus reasoning systems will help democratize information, and increase LLM safety by increasing our ability to process LLM generations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Yes, our motivation and task discussion (e.g. contradiction task to cross-refer scientific abstract claims) generally tie our method to exciting positive societal applications. We don't find direct negative applications to be applicable to this work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work doesn't involve the usage of any high-risk data, all the data we use is already open-source and adequately credited and cited.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, we explicitly cite original owners of all data we use to construct our experimental settings, model providers, and so forth. We stick to using open source artifacts.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, this paper introduces several new data artifacts (both training and evaluation) as well as tasks, which we provide ample code and instruction on how to reproduce.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.