Diversity Enhances an LLM's Performance in RAG and Long-context Tasks

Anonymous ACL submission

Abstract

The rapid advancements in large language models (LLMs) have highlighted the challenge of context window limitations, primarily due to the quadratic time complexity of the self-attention mechanism ($O(N^2)$, where N denotes the context window length). This constraint impacts tasks such as retrieval-augmented generation (RAG) in question answering (Q&A) and long context summarization. A common approach involves selecting content with the highest similarity to the query; however, this often leads to redundancy and the exclusion of diverse yet relevant information. Building on principles from Maximal Marginal Relevance (MMR) and Farthest Point Sampling (FPS), we integrate diversity into the content selection process. Our findings reveal that incorporating diversity substantially increases the recall of selecting relevant sentences or chunks before LLM-based Q&A and summarization. These results highlight the importance of maintaining diversity in future LLM applications to further improve summarization and Q&A outcomes.

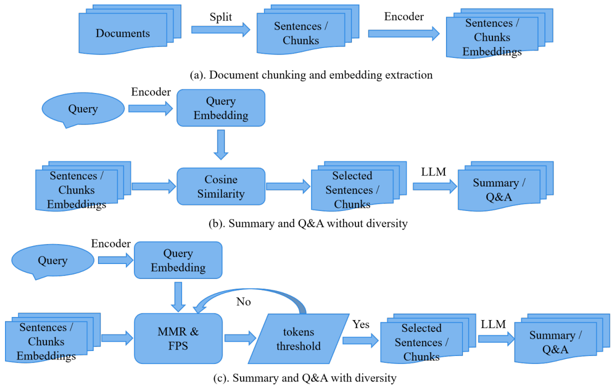

Figure 1: For both Q&A and summarization tasks, the initial dataset is divided into sentences or chunks, and corresponding embeddings are extracted. In a traditional pipeline, query embeddings are generated and used to select relevant materials to LLMs for downstream tasks. In contrast, methods like MMR and FPS incorporate diversity in a greedy manner when selecting relevant sentences. This approach increases the likelihood of including the correct answer within the chosen sentences or chunks.

1 Introduction

The remarkable success of Transformer models (Vaswani et al., 2023), BERT (Devlin et al., 2019), and GPT (OpenAI et al., 2024) can be largely attributed to their robust self-attention mechanisms. However, the self-attention module's quadratic time complexity, $O(N^2)$, where N represents the context window length, has imposed limitations on the size of the context window.

Recent advances in LLMs have partially addressed this constraint. For instance, GPT-3.5 demonstrates the capability to process context windows of up to 16,385 tokens, while GPT-4 extends this capacity to an impressive 128,000 tokens. Despite these notable improvements, the challenge of processing even longer sequences remains a critical area of research for several compelling reasons. First, many real-world applications, such as question-answering systems operating on extensive datasets, cannot accommodate entire document collections within the LLM's context window. This limitation has led to the development of Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2021), which selectively retrieve and process relevant text segments for specific queries. Second, while current context window sizes may suffice for conventional Natural Language Processing (NLP) tasks, they prove inadequate for high-frequency signal processing applications. For example, audio processing and medical vibrational signal analysis often require handling data streams with sampling rates reaching one million samples per second, far exceeding current context window capabilities (Gu and Dao, 2024). Furthermore, empirical studies have revealed a concerning trend: LLM performance tends to degrade as input lengths approach the maximum context window capacity, highlighting the need for more robust solutions to long-sequence processing (Nvidia et al., 2024).

Various strategies have been devised to address the limited context window issue in LLMs. Longformer (Beltagy et al., 2020) applies attention to immediate local neighbors, reducing the time complexity from $O(N^2)$ to $O(NM)$, with M representing the considered neighbors. This approach, however, necessitates significant alterations to the attention mechanism, which is not commonly adopted in contemporary LLMs such as GPT(OpenAI et al., 2024), Llama(Touvron et al., 2023), and Gemini(Team et al., 2024). An alternative strategy is to expand the context window at inference (Jin et al., 2024). Although this can mitigate the modification during the training process, it still demands changes to the attention architecture during the inference time, which is not accessible for close-source models like GPT.

Several strategies have been proposed to address the limited context window in LLM from the training perspective. The Longformer model (Beltagy et al., 2020) employs attention mechanisms focused on immediate local neighbors, reducing the time complexity from $O(N^2)$ to $O(NM)$, where M denotes the number of neighbors considered. However, this method requires substantial modifications to the attention mechanism, which are not widely adopted by contemporary LLMs such as GPT (OpenAI et al., 2024), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2024). Another approach involves extending the context window during inference (Jin et al., 2024). While this mitigates the need for training-time modifications, it necessitates changes to the attention mechanism at inference and full access to the model architecture—an obstacle for closed-source models like GPT.

Previous methods primarily focus on modifying LLMs to increase their context window. However, a more straightforward approach is to first select the most relevant documents while ensuring they fit within the LLM's context window. For a given query, multiple documents are split into smaller chunks or sentences. The embeddings for both the query and the split documents are then computed. Similarity metrics, such as cosine similarity or Euclidean distance, are subsequently used to identify the most relevant sentences.

However, relying solely on the similarity between a query and segmented documents can result in overlooking critical information due to excessive focus on similar content. Previous studies have introduced greedy algorithms, such as MMR (Carbonell and Goldstein, 1998) and FPS (Qi et al., 2017), to improve diversity during the selection process. Related work introduced Hypothetical Document Embedding (HyDE) and LLM reranking to enhance diversity in Q&A tasks, claiming their method outperforms MMR (Eibich et al., 2024; Pickett et al., 2024). However, these studies did not address the recall of relevant documents prior to LLM generation, which is more pertinent to diversity considerations. Additionally, they did not explore various hyperparameters within MMR. Then, they have not explored the impact of reordering of selected sentences or chunks on the downstream tasks. In this paper, we aim to address this gap by conducting experiments to demonstrate the significance of diversity in long context summarization and RAG-based Q&A tasks at multiple levels: sentence-level for single documents, chunk-level across entire datasets, and sentence-level in summarization.

The contributions of this paper are summarized as follows:

1. We demonstrate the benefits of diversity using MMR and FPS with proper hyperparameters, i.e., α and w on downstream tasks, including Q&A and summarization.

2. We discover that MMR achieves slightly better recall than FPS while maintaining significantly lower latency.

3. We prove the ordering selected sentences within the original document and ordering selected chunks based on the scores has the best downstream performances.

2 Methodology

In this section, we will start with a brief introduction of MMR and FPS to consider diversity during the search process. Then, the integration with LLM will be discussed.

2.1 MMR and FPS for Diversity

MMR The concept of MMR involves selecting a subset S from a large dataset T (Carbonell and Goldstein, 1998). MMR uses a greedy algorithm that starts with the selected set S being empty and the remaining set R being the entire dataset T. In each iteration, an element is chosen based on a locally optimal selection process, as defined in Eq. 1. The parameter α balances the trade-off between rewards and diversity. Let r_i denote the reward of the i-th item, and $cos(i, j)$ represent the cosine similarity between the i-th and j-th items in the selected subset. W is a subset of S that includes the most recently selected examples, reducing the emphasis on earlier selections. For example, if $w = 10$, W consists of the last 10 selected samples from S, while all previously selected examples are excluded from diversity considerations. The objective of MMR is to maximize rewards while ensuring sufficient diversity among the selected items. This iterative process continues until a termination criterion, such as reaching a predefined maximum number of tokens, is met.

$$\underset{i \in R}{\operatorname{argmax}} \left[\alpha \cdot r_i - (1 - \alpha) \cdot \max_{j \in W} \cos(i, j) \right] \quad (1)$$

FPS The concept of FPS originates from the field of 3D computer vision (Qi et al., 2017). Its primary goal lies in selecting a diverse set of points from a given point cloud, which aids in hierarchical feature extraction for downstream applications. The process begins with a randomly selected initial point. In each subsequent iteration, a new point is chosen based on its distance from all previously selected points. When comparing FPS to MMR, we find that both are greedy methods that promote diversity by selecting points that differ from those chosen. However, FPS does not incorporate the concepts of a context window or reward. If we modify FPS to include these elements, the modified FPS will be equivalent to MMR, with the key difference being that MMR uses cosine similarity, while FPS relies on Euclidean distance for measuring similarity.

$$\underset{i \in R}{\operatorname{argmax}} \left[\max_{j \in S} \operatorname{dist}(i, j) \right] \quad (2)$$

2.2 Combine MMR and FPS with LLM for Diversity on Q&A and Summarization

Extending MMR and FPS techniques for LLMs in tasks such as Q&A and summarization is relatively straightforward as shown in Fig. 1. These techniques employ a greedy approach to iteratively balance the similarity of selected sentences or chunks to the query with the diversity among the selected sentences or chunks. This method enhances the likelihood of selecting the most relevant sentences or chunks for LLMs in downstream tasks. Lastly, inspired by (Liu et al., 2023), a heuristic rearrangement scheme is implemented to enhance the likelihood of identifying the correct answer from the retrieved documents.

Q&A To evaluate the ability of LLMs on accurately extracting the correct answer, a query, a document, and a corresponding answer are initially provided. Documents are pre-processed by dividing them into sentences or chunks, and their embeddings are extracted beforehand. Both the query and the segmented documents are processed using encoder-only models to generate embeddings. In MMR, similarity is measured using the cosine angle, whereas in FPS, Euclidean distance is used to assess similarity. For benchmarking Q&A performance, two metrics should be evaluated:

1. Pre-LLM recall: whether the answer exists in the selected content before being sent to the LLM.

2. Post-LLM recall: whether the answer appears in the LLM's output.

If the first metric shows significant improvement, the benefit of diversity becomes evident. Otherwise, the advantage of diversity may be limited. If the first metric improves while the second metric does not, it indicates that the performance of downstream tasks may be constrained by the capabilities of the LLM (Liu et al., 2023).

Summarization In summarization tasks, datasets typically consist of a document paired with a corresponding golden summary created by experts. When no specific query is provided, the process begins by dividing the document into manageable chunks. Encoder-only models are employed to generate embeddings for these chunks, and the mean of these embeddings is used to represent the query embedding. Following this, the same methodology as in the previous Q&A task is applied to extract content that optimizes both reward and diversity.

The selected chunks are ordered to align with their original sequence in the document. These ordered chunks are sent to the LLM for summarization. We recognize that evaluating the extracted content before it is submitted to the LLM for summarization may not be particularly meaningful. Instead, we assess the quality of the LLM-generated summary by comparing it to the golden summary

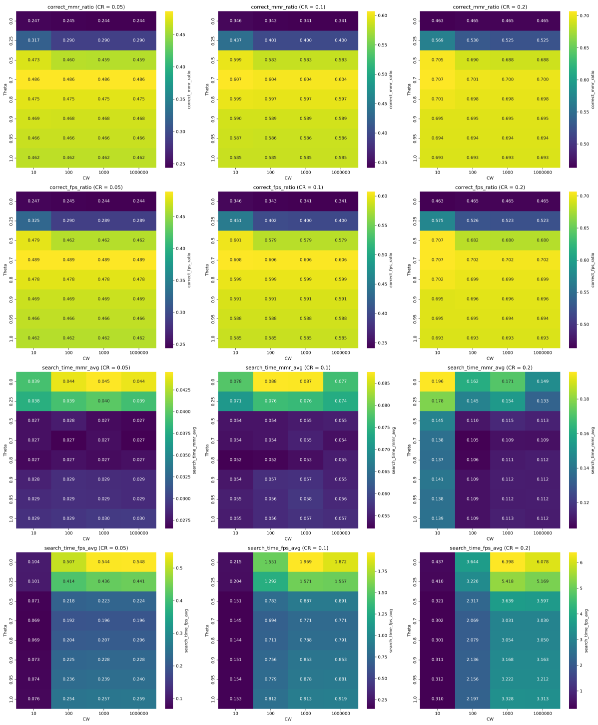

Figure 2: The impact of different hyperparameters: α, w, c_r on the recall of the Natural Question dataset of single document Q&A. The first and second subfigures illustrate the recall ratios of answers contained in the selected documents for SB+MMR and SB+FPS. When the weight parameter $w = 1$, they are equivalent to SB. From the results, we can conclude that both SB+MMR and SB+FPS outperform SB. The last two subfigures display the latency of SB+MMR and SB+FPS. SB+FPS shows slightly worse performances than SB+MMR, and the latency of SB+MMR is significantly lower, especially when the context window is very long. Considering these two aspects, SB+MMR is more suitable for practical use compared to SB+FPS.

using metrics such as ROUGE (Lin, 2004) or LLM-as-a-judge (Hsu et al., 2024).

3 Experiments

The experiments conducted in this paper focus on three main topics: 1. Single Document Question Answering (Q&A), 2. Multiple Documents Question Answering (Q&A), and 3. Single Document Summarization.

For Single Document Q&A, the goal is to choose the correct answer from a set of candidate sentences within a single document. In Multiple Document Q&A, all documents in the dataset are firstly divided into chunks and then combined, and a query is used to find the correct answers across the entire dataset. Because the dataset size is too large, approximation methods are used to enhance efficiency and speed. Specifically, two metrics are evaluated: 1. recall of the correct answer in the extracted document, and 2. recall of the correct answer in the LLM response. The benefit of diversity is primarily reflected in the improvement of the first metric, while performance improvements in Q&A and summarization are mainly indicated by the second metric.

For summarization, various hyperparameters are considered in the optimization process:

1. The weight balance between reward and diversity, denoted as α,
2. The context window size, w,
3. The compression ratio, c_r, or the maximum number of selected tokens, T_{max}.

3.1 Single Document Q&A

For single document Q&A, three datasets are included: 1. Natural Question (Kwiatkowski et al., 2019), 2. Trival QA (Joshi et al., 2017) and 3. Narrative QA (Kočiský et al., 2017). For each dataset, it is composed of thousands of (query, document, answer) pairs where the answer exists within the document and answers the query. For each document, we split it into sentence using Spacy package (Honnibal et al., 2020). SentenceBERT (SB) is utilized as the encoder to extract embeddings from sentences in different experiments (Reimers and Gurevych, 2019). Then, different compression ratios, i.e., $c_r = 0.05, 0.1, 0.2$ are utilized. Here, $c_r = 0.05$ represents that the fraction of the number of selected tokens over the total number of tokens should be 0.05, i.e., the termination condition when 0.05 of all tokens are selected for answering the query. As for α and w, different hyperparameters are tested in a two-level iteration. The first coarse-level iteration utilizes α from [0, 0.25, 0.5, 0.7, 0.8, 0.9, 0.95, 1] and w from [0, 10, 100, 1000, 1000000]. Then, the best performing from the first coarse-level iteration is selected. For the second granular-level iteration, it further divides the neighbors of the best performing first coarse-level hyperparameters and select the ones that have the best performance. An example of the best performing hyperparameters in the Natural Question dataset in the coarse level can be found in Fig. 2. In particular, for Natural question, $\alpha = 0.55$ and $w = 1$ is the best for MMR and $\alpha = 0.5$ and $w = 1$ is the best for FPS. For Narrative Q&A, $\alpha = 0.55$ and $w = 5$ is the best for both MMR and FPS. For

Q&A	Natural Question			Trival Q&A			Narrative Q&A		
	c_r=0.05	c_r=0.1	c_r=0.2	c_r=0.05	c_r=0.1	c_r=0.2	c_r=0.05	c_r=0.1	c_r=0.2
SB	46.28	58.60	69.41	63.44	71.64	78.33	18.60	21.04	25.61
SB+MMR	50.43	63.18	**72.47**	**65.29**	**74.02**	**80.47**	20.88	24.09	**27.59**
SB+FPS	**50.88**	**63.23**	72.33	65.25	73.18	80.07	**21.34**	**24.24**	27.44

Table 1: This table compares the performance of SB+MMR and SB+FPS against SB across three different datasets and three compression ratios, focusing on the recall of the correct answer within the selected documents.

Q&A	Natural Question			Trival Q&A			Narrative Q&A		
	c_r=0.05	c_r=0.1	c_r=0.2	c_r=0.05	c_r=0.1	c_r=0.2	c_r=0.05	c_r=0.1	c_r=0.2
SB (index sort)	40.44	51.74	64.25	75.66	75.95	76.25	15.40	17.98	18.60
SB (sort)	40.25	50.81	60.61	75.76	76.15	76.38	13.72	15.55	17.07
SB (1:1)	40.25	51.55	60.24	75.29	76.33	76.47	14.18	16.31	17.99
SB (2:1)	41.28	50.99	61.08	75.31	76.33	76.87	14.18	16.46	17.84
SB (3:1)	39.41	51.09	60.05	75.56	76.13	76.57	14.63	16.01	17.23
SB+MMR (index sort)	45.76	57.25	**67.90**	75.73	76.20	76.72	**18.60**	**19.05**	**20.27**
SB+MMR (sort)	44.45	55.19	64.35	76.20	**76.85**	77.00	16.31	16.46	17.38
SB+MMR (1:1)	45.48	55.57	63.60	76.23	76.77	**77.34**	16.46	16.62	16.92
SB+MMR (2:1)	45.20	55.29	63.32	75.90	76.20	76.65	15.09	15.55	16.01
SB+MMR (3:1)	43.80	55.85	63.51	76.05	76.72	76.92	16.16	16.92	16.31
SB+FPS (index sort)	**46.79**	**59.12**	67.71	75.93	76.13	76.60	17.68	17.68	19.05
SB+FPS (sort)	45.76	57.25	63.32	75.83	76.45	76.87	16.31	16.62	17.23
SB+FPS (1:1)	46.14	57.90	62.76	75.78	76.35	77.09	16.62	17.07	16.77
SB+FPS (2:1)	45.76	56.13	62.20	75.88	76.23	77.02	15.85	16.62	16.46
SB+FPS (3:1)	44.27	55.10	63.69	**76.40**	76.77	76.95	15.70	16.46	16.01

Table 2: This table compares the performance of SB+MMR and SB+FPS against SB across three different datasets and three compression ratios, focusing on the recall of the correct answer within the LLM responses.

Trival Q&A, $\alpha = 0.6$ and $w = 3$ is the best for MMR and $\alpha = 0.7$ and $w = 1$ is the best for FPS.

Based on the results across various datasets, we can assert that diversity significantly enhances the recall of the correct answer within the selected document, as demonstrated in Table 1, showing an improvement of 2% to 5%. When the extracted sentences are summarized by GPT4 using the prompt shown in Figure 3, the advantages of SB+MMR and SB+FPS over SB alone remain evident, as shown in Table 2. Additionally, we observe that the performance of Trivial Q&A after LLM is better than the retrieved sentences, with a consistent result of approximately 76%. This suggests that the performance is largely influenced by the LLM, possibly due to pretraining on Trivial Q&A, even when the retrieved documents are provided. FPS, using distance as the evaluation metric, performs slightly worse than MMR, which uses cosine similarity. Moreover, MMR is faster than FPS because computing cosine similarity is quicker than Euclidean distance in Python, especially as the compression ratio increases, as shown in Figure 2. This conclusion generally holds true across different datasets. The speed advantage of MMR becomes more critical as the number of candidates increases with the dataset size. Consequently, MMR will be used in the multiple document comparison in the next section.

Inspired by the paper "Lost in the Middle" (Liu et al., 2023), we sorted the selected sentences by different methods. The term "index sort" refers to sorting the sentences in their original order within the document. In comparison, "SB (m:n)" refers to allocating the first selected m sentences with highest scores at the beginning, the next n sentences with highest scores at the end, and then another m sentences at the beginning, continuing this pattern until all sentences are allocated. Specifically, "SB (sort)" is equivalent to "SB (1:0)" and does not alter the sequence of selected sentences. As shown in Table 2, SB (index sort) performs best because the original sequential information of the selected sentences in the document, despite missing some internal information, makes the most sense for GPT-4 in downstream tasks.

3.2 Mutiple Documents Q&A

For multiple documents Q&A, the same three datasets are utilized. In these datasets, the number of documents and the length of documents are relatively long, making it impractical to split each document into sentences. Instead, we follow the general framework of RAG to split each document into chunks of 512 tokens, with an overlapping ratio of 0.5 (i.e., 256 tokens) between any two ad-

Natural Question	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5	70.7	69.4	68.5	66.2	64.2	57.4
E5+MMR	**71.5**	**71.5**	**69.8**	**67.2**	**65.1**	**57.8**
Narrative Q&A	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5	13.42	10.06	6.7	4.88	4.88	4.57
E5+MMR	**22.56**	**20.43**	**15.85**	**14.94**	**12.20**	**7.01**
Trival Q&A	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5	84.62	81.15	74.01	70.24	65.08	56.55
E5+MMR	**88.99**	**85.81**	**82.24**	**78.57**	**74.01**	**65.08**

Table 3: This table compares the performance of E5+MMR against E5 across three different datasets, focusing on the recall of the correct answer within the selected documents.

Natural Question	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5 (index sort)	45.7	50.5	52.7	45.9	49.2	45.4
E5 (sort)	55.2	56.6	54.6	51.5	50.5	47.8
E5 (1:1)	53.5	55.9	55.8	50.5	50.9	46.9
E5 (2:1)	54.5	57.5	56.4	51	50.5	47.1
E5 (3:1)	54.7	57	54.8	51.3	49.8	47.4
E5+MMR (index sort)	46.2	50.3	53.2	47.6	50.9	48.6
E5+MMR (sort)	56.4	**57.9**	**57**	51.3	51.4	47.7
E5+MMR (1:1)	55	57.2	55.9	50.8	50.7	47.8
E5+MMR (2:1)	**57.2**	55.3	56.2	50.7	52.2	**48.7**
E5+MMR (3:1)	56.3	56.4	55.6	**51.6**	**52.3**	47.4

Table 4: This table compares the performance of E5+MMR against E5 on Natural Question, focusing on the recall of the correct answer within the LLM responses.

Narrative Q&A	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5 (index sort)	10.37	9.15	8.54	5.18	4.57	4.88
E5 (sort)	10.67	10.59	8.23	6.1	4.57	5.18
E5 (1:1)	9.76	9.45	7.93	4.88	4.88	5.18
E5 (2:1)	10.06	9.15	7.93	5.18	3.96	5.18
E5 (3:1)	10.98	9.45	7.93	5.79	4.57	4.88
E5+MMR (index sort)	10.67	10.37	**11.28**	4.88	6.1	4.88
E5+MMR (sort)	**12.8**	10.67	10.37	6.1	6.1	4.27
E5+MMR (1:1)	11.89	10.06	**11.28**	6.4	6.71	4.88
E5+MMR (2:1)	11.59	10.67	10.98	6.4	5.79	**5.49**
E5+MMR (3:1)	11.89	**10.98**	10.06	**6.7**	**7.01**	4.88

Table 5: This table compares the performance of E5+MMR against E5 on Narrative Q&A, focusing on the recall of the correct answer within the LLM responses.

Trival Q&A	GPT4			GPT3.5		
	T_{max}=120k	T_{max}=50k	T_{max}=20k	T_{max}=10k	T_{max}=5k	T_{max}=2k
E5 (index sort)	74.21	73.21	73.12	65.18	64.19	64.68
E5 (sort)	73.81	73.91	72.42	63.59	65.08	65.57
E5 (1:1)	74.4	73.12	72.92	64.29	64.29	64.98
E5 (2:1)	73.81	73.81	72.72	64.09	64.58	64.29
E5 (3:1)	73.31	74.11	72.72	63.99	64.38	64.78
E5+MMR (index sort)	74.7	74.9	**73.51**	66.47	**66.87**	64.88
E5+MMR (sort)	74.9	74.8	73.12	64.29	65.57	**66.07**
E5+MMR (1:1)	**75.2**	**75.5**	73.31	65.38	65.38	65.08
E5+MMR (2:1)	74.6	74.11	72.62	64.88	65.77	65.38
E5+MMR (3:1)	74.7	74.31	73.02	**65.67**	65.48	64.98

Table 6: This table compares the performance of E5+MMR against E5 on Trival Q&A, focusing on the recall of the correct answer within the LLM responses.

jacent chunks. To extract embeddings from these chunks and adhere to the standard pipeline of RAG, we apply the E5 model (Wang et al., 2024). After applying the chunking strategy, the number of chunks can still reach nearly 1 million, which is impractical for exact search. To facilitate approximate search, principal component analysis (PCA) (Maćkiewicz and Ratajczak, 1993) is first applied to reduce the dimensionality of the embeddings, followed by clustering (Bishop and Nasrabadi, 2006) to ensure the average number of chunks is less than 10k. Unlike single document Q&A, we set the maximum number of tokens rather than the compression ratio as the threshold for the maximum number of tokens selected. Specifically, T_{max} is set to 2k, 5k, or 10k for GPT-3.5 and 20k, 50k, or 120k for GPT-4. Other settings remain the same. Different hyperparameters for α and w are tested. For the Natural Questions dataset, $\alpha = 0.9$ and $w = 5$ yield the best results for GPT-3.5, while $\alpha = 0.7$ and $w = 5$ are optimal for GPT-4 in MMR. For Narrative Q&A, $\alpha = 0.8$ and $w = 30$ are best for GPT-3.5, and $\alpha = 0.7$ and $w = 300$ are best for GPT-4 in MMR. For Trivia Q&A, $\alpha = 0.7$ and $w = 20$ are best for GPT-3.5, and $\alpha = 0.8$ and $w = 300$ are best for GPT-4 in MMR. From the results, we observe that the optimal values for α and w are generally larger for multiple document Q&A compared to single document Q&A.

When evaluating the performance of multiple-document Q&A systems, we observe a pattern similar to that of single-document Q&A. Specifically, the E5+MMR method shows a significant improvement over E5 in recall of the answers in retrieved documents, as demonstrated in Table 3, with a margin exceeding 10%. Additionally, E5+MMR outperforms E5 for post-LLM recall as shown in Tables 4, 5, and 6. However, future research should prioritize enhancing the LLM's ability to utilize the retrieved documents effectively, rather than merely focusing on retrieving more accurate documents, as the LLM itself is the bottleneck. This observation is further corroborated in Trivial Q&A, where the results consistently achieve 64% accuracy for GPT3.5 and 76% for GPT4, irrespective of the retrieved document. Last, unlike single-document Q&A, placing important chunks at the beginning and ending positions of the prompt can provide benefits, particularly in Natural Question scenarios, as shown in Table 4, which can lead to a 10% improvement. This finding aligns with the conclusions of the paper "Lost in the Middle". The most relevant chunks to the query should be positioned either at the beginning or the end of the prompt.

3.3 Sentence and Chunk Splitter Comparison on SquAD

For the comparison between sentence and chunk splitters on multiple documents Q&A, only the SQuAD dataset will be considered. The dataset sizes for Natural Questions, TriviaQA, and NarrativeQA are too large, making sentence-level experiments difficult. For the sentence-level splitter, Spacy is used. For the chunk-level splitter, a threshold of 256 or 512 tokens with 50% overlap between adjacent chunks is applied. All segmented sentences or chunks are mixed, reduced in dimension through PCA, and clustered for downstream tasks. Similar to previous experiments, T_{max} is set to 2k, 5k, or 10k for GPT3.5. For the sentence-level splitter, the best parameters are $\alpha = 0.25$ and $w = 1000$. For the 256-token chunk-level splitter, the best parameters are $\alpha = 0.5$ and $w = 300$. For the 512-token chunk-level splitter, the best parameters are $\alpha = 0.3$ and $w = 300$. The results are consistent with previous findings. SB/E5+MMR significantly outperforms SB/E5, as shown in Table 7, with a 10% increase in recall of the correct answer within the selected documents. This recall increment of SB/E5+MMR over SB/E5 still exists in the LLM response, as shown in Table 8. "Index sort" generally performs better for sentence-level splitting, while sorting based on score is usually beneficial for chunking. A new takeaway is that chunk-level performance is better than sentence-level, with even better results for larger chunk sizes.

3.4 Single Document Summarization

For single document summarization, we include two datasets: the gov report (Huang et al., 2021) and legal documents (Shukla et al., 2022). We utilize GPT3.5 for summarization. To achieve this, we filter examples that are less than 15k tokens and then apply MMR to select sentences within each document until it reaches the predetermined threshold of 8k tokens. For the 1. gov report, the best parameters are $\alpha = 0.9$ and $w = 10$. For the legal documents, the best parameters are $\alpha = 0.925$ and $w = 300$. After selecting and ordering the selected sentences based on their original sequence, they are sent to GPT3.5 to generate the final summary using a specific prompt in Figure. 4. For both datasets, expert-written golden summaries are provided for each document. We evaluate the quality of generated summary using the ROUGE score by comparing with the golden summary. In addition,

SquAD	Sentence			Chunk size: 256			Chunk size: 512		
	10k	5k	2k	10k	5k	2k	10k	5k	2k
SB/E5	86.8	83.7	78.5	95	92.7	86.3	96.7	94.3	86.6
SB/E5+MMR	**90.1**	**89.4**	**86.6**	**97**	**96.6**	**95.4**	**99**	**97.8**	**96.7**

Table 7: This table compares the performances of sentence splitter and chunk splitter of size 256 and 512 on SquAD, focusing on the recall of the correct answer within the selected documents.

SquAD	Sentence			Chunk size: 256			Chunk size: 512		
	10k	5k	2k	10k	5k	2k	10k	5k	2k
SB/E5 (index sort)	70.4	73	71.7	78.2	81.1	79.2	80.2	82.9	78.2
SB/E5 (sort)	71.3	70.1	67.9	79.9	80	76.6	82.9	82.2	76.5
SB/E5 (1:1)	72	71.8	69.5	80.3	80.4	77.5	82.7	82	77.4
SB/E5 (2:1)	71.9	71.7	69.4	80.4	80.1	77.4	82.9	81.9	77.5
SB/E5 (3:1)	71.8	71	68.7	81.4	79.9	77.7	82.7	82.2	77.3
SB/E5+MMR (index sort)	68.7	**74.2**	**75**	76.5	81.4	83	76.1	84.2	84.9
SB/E5+MMR (sort)	71.5	71.8	72.5	82	81.9	82.6	84.1	**84.5**	84.7
SB/E5+MMR (1:1)	**72.5**	73.6	72.9	81.3	82.2	81.7	82.3	82.9	84.6
SB/E5+MMR (2:1)	71.9	73.1	73.2	**81.8**	**82.5**	**83.1**	**84.5**	**84.5**	**85**
SB/E5+MMR (3:1)	71.7	73.4	71.9	81.4	**82.5**	82.9	83.4	84.3	84.5

Table 8: This table compares the performances of sentence splitter and chunk splitter of size 256 and 512 on SquAD, focusing on the recall of the correct answer within the LLM responses.

Summarization Datasets	gov_report		legal	
	Rouge	GPT4 WR	Rouge	GPT4 WR
SB	17.7	24.24	11.3	35.82
SB+MMR	**18**	$\frac{72.65+77.86}{2} = \mathbf{75.26}$	**11.8**	$\frac{71.64+56.72}{2} = \mathbf{64.18}$

Table 9: This table compares the performance of SB+MMR against SB on gov_report and legal, using ROUGE and LLM-as-a-judge.

summaries by SB and SB+MMR are compared using LLM-as-a-Judge through GPT4. To address the position bias problem, we switch the sequences of the two summaries in two runs and average the win rate (WR). Our experiments reveal that diversity improves summary quality, as indicated by increased ROUGE scores and a higher LLM-as-a-Judge WR. Additionally, experiments on our internal data show that diversity is particularly beneficial for long emails, articles, and logs, where redundancy is a significant issue due to repetitive content, greetings, and long URLs. Diversity avoid overestimating information similar to the query.

4 Conclusion

This study proves the benefits of diversity through MMR and FPS to LLM performances on Q&A and summarization. From the retrival viewpoint, the recall is greatly improved both for sentence and chunk-level splitter, especially when α and w are properly selected. This recall rate increment is maintained after LLM generation. However, future research should pay more attention to improve the LLM's capability to find answers from the retrieved documents. MMR shows slightly better performances compared with FPS, and its latency property is much better, which greatly increase the potential of usage in application. For sentence-level splitter, arranging the selected sentences in their original sequence is usually beneficial and for chunk-level splitter, putting more important chunks at the beginning and ending positions are beneficial. Lastly, given a multiple document Q&A like SquAD, chunk-level splitter usually has a better performance compared with sentence-level splitter. Lastly, these conclusion on Q&A can be extended to summarization task.

5 Limitation

There are several limitation on this works. To begin with, we only work on English dataset, while multilingual datasets should be tested to prove the importance of diversity on other language. In addition, this work focuses on research dataset while more work is supposed to be conducted on industrial datasets. Lastly, for extremely large dataset, more engineering work on parallelization like tree structures should be conducted to reduce latency.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Matouš Eibich, Shivay Nagpal, and Alexander Fred-Ojala. 2024. Aragog: Advanced rag output grading. *Preprint*, arXiv:2404.01037.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Aliyah R. Hsu, James Zhu, Zhichao Wang, Bin Bi, Shubham Mehrotra, Shiva K. Pentyala, Katherine Tan, Xiang-Bo Mao, Roshanak Omrani, Sougata Chaudhuri, Regunathan Radhakrishnan, Sitaram Asur, Claire Na Cheng, and Bin Yu. 2024. Rate, explain and cite (rec): Enhanced explanation and attribution in automatic evaluation by large language models. *Preprint*, arXiv:2411.02448.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *Preprint*, arXiv:2104.02112.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *Preprint*, arXiv:2401.01325.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Preprint*, arXiv:1712.07040.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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 Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. 2024. Nemotron-4 340b technical report. *Preprint*, arXiv:2406.11704.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Marc Pickett, Jeremy Hartman, Ayan Kumar Bhowmick, Raquib ul Alam, and Aditya Vempaty. 2024. Better rag using relevant information gain. *Preprint*, arXiv:2407.12101.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Preprint*, arXiv:1706.02413.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. *Preprint*, arXiv:2210.07544.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Jun-

hyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Siddharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Al-

lica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

A Example Appendix

A.1 Prompts for Q&A and Summarization

Here are the prompts for Q&A in Figure. 3 and summarization in Figure. 4.

```
You are tasked with answering a query based on the provided context.
Respond concisely by directly citing a relevant portion of the original context.

query:
###
{query}
###

context:
###
{context}
###

answer (exactly copy from the context):
```

Figure 3: Prompts for Q&A

```
Please summarize based on the following context.
The summary should incorporate both qualitative and quantitative information.
The qualitative section should highlight central themes, emerging trends, and critical elements.
Meanwhile, the quantitative section should present supporting statistics and numerical data relevant to the summary.

Context:
###
{context}
###

Summary:
```

Figure 4: Prompts for Summarization