

Machine Learning Research Context + Contribution

In the summer of 2023, a few close friends and I came together to research knowledge feeding into Large Language Models (LLMs) in the hopes of making them more efficient. The research was done online using virtual code editors and Jupyter notebooks with added GPU computation. We have been working on this project since the summer of 2023 and have been trying to publish it for a few months now. We are currently in the conditional approval stage of the journal, ACM Transactions on Intelligent Systems and Technology, with a likely publication after their internal processes. Regarding the workload of the project, I did about 40%.

We didn't have any mentors. I asked a few people for advice regarding methodology questions, but there were no external contributions to the research. The three of us worked as a team.

First Author: Santosh Kumar

Second Author: Me

Third Author: Supriya Devadutta


Fourth Author: Karthikeyan S (mainly helped with the publication process)

DOI Link (published pre-print version on arXiv)

<https://doi.org/10.48550/arXiv.2502.05233>

Screenshot of Conditional Acceptance email from ACM TIST journal

[ACM TIST], TIST-2024-12-0933, Minor Revision External Inbox x Print Share

 **Transactions on Intelligent Systems and Technology** <onbehalf@manuscriptcentral.com> Dec 17, 2024, 10:16 PM Star Reply More

to me, sthanikamsanthosh1994, tsharma, supriyadevudutta.ml8, gita.delsing, yogalakshmi.m ▾

17-Dec-2024

Dear Mr. Rishi Gottimukkala:

Manuscript TIST-2024-12-0933 entitled "Efficient Knowledge Feeding to Language Models: A Novel Integrated Encoder-Decoder Architecture" which you submitted to the Transactions on Intelligent Systems and Technology, has been reviewed. The comments of the reviewer(s) are included at the bottom of this letter.

I have received the reviews and the recommendation from the Associate Editor on your paper and I conclude that the paper can be conditionally accepted in its present form, but requires a minor revision and re-review.

If you decide that you would like to make minor revisions, please read the reviews carefully and upload your revised manuscript within three months of today's date (due date: 17-Mar-2025).

Please follow the procedure below for resubmitting your manuscript.

- (1) Once you have revised your manuscript, go to <https://mc.manuscriptcentral.com/tist> and login to your Author Center.
- (2) Click on "Manuscripts with Decisions," and then click on "Create a Resubmission" located next to the manuscript number.
- (3) Follow the steps for resubmitting your manuscript.

In addition to your revised manuscript, please also include a document entitled "Responses to Reviewers' Comments" that explains how individual comments and suggestions of the Reviewers were incorporated into the revised manuscript. You can also include a document entitled "Summary of Differences" that describes the differences from the previous version.

Thank you for submission to the Transactions on Intelligent Systems and Technology and I look forward to receiving your revision.

Sincerely,

Huan Liu
Editor in Chief, Transactions on Intelligent Systems and Technology
Homepage: <http://www.public.asu.edu/~huanliu/>

Efficient Knowledge Feeding to Language Models: A Novel Integrated Encoder-Decoder Architecture

S SANTHOSH KUMAR
RISHI GOTTIMUKKALA
SUPRIYA DEVIDUTTA
KARTHIKEYAN S

This paper introduces a novel approach to efficiently feeding knowledge to language models (LLMs) during prediction by integrating retrieval and generation processes within a unified framework. While the Retrieval-Augmented Generation (RAG) model addresses gaps in LLMs' training data and knowledge limits, it is hindered by token limit restrictions and dependency on the retrieval system's accuracy. Our proposed architecture incorporates in-context vectors (ICV) to overcome these challenges. ICV recasts in-context learning by using latent embeddings of LLMs to create a vector that captures essential task information. This vector is then used to shift the latent states of the LLM, enhancing the generation process without adding demonstration examples to the prompt. ICV directly integrates information into the model, enabling it to process this information more effectively. Our extensive experimental evaluation demonstrates that ICV outperforms standard in-context learning and fine-tuning across question answering, information retrieval, and other tasks. This approach mitigates the limitations of current RAG models and offers a more robust solution for handling extensive and diverse datasets. Despite leveraging a fraction of the parameters, our ICV-enhanced model achieves competitive performance against models like LLaMA-3, Gemma, and Phi-3 significantly reducing computational cost and memory requirements. ICV reduces prompt length, is easy to control, surpasses token limitations, and is computationally efficient compared to fine-tuning.

CCS Concepts: • **Computing methodologies** → **Information extraction; Natural language generation; Knowledge representation and reasoning; Software and its engineering; Information systems** → *Query representation; Query reformulation; Language models; Top-k retrieval in databases; Novelty in information retrieval; Question answering; Information extraction; Relevance assessment; Retrieval effectiveness; Retrieval efficiency; Presentation of retrieval results; Environment-specific retrieval;*

Additional Key Words and Phrases: Knowledge Feeding, LLM, ICV, RAG, Encoder-Decoder

ACM Reference Format:

S Santhosh Kumar, Rishi Gottimukkala, Supriya Devidutta, and Karthikeyan S. 2024. Efficient Knowledge Feeding to Language Models: A Novel Integrated Encoder-Decoder Architecture. *ACM Trans. Intell. Syst. Technol.* XX, X, Article XXX (X 2024), 15 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The advent of large language models (LLMs) such as GPT-3, GPT-4, and Llama has revolutionized the field of natural language processing [4, 20, 24], enabling impressive advancements in applications ranging from natural language understanding to sophisticated content generation [21, 25, 37]. These models, trained on vast amounts of text data, possess the ability to generate human-like responses and perform complex linguistic tasks. However, despite their remarkable capabilities,

Authors' Contact Information: S Santhosh Kumar, ; Rishi Gottimukkala, ; Supriya Devidutta, ; Karthikeyan S, .

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2157-6912/2024/X-ARTXXX
<https://doi.org/XXXXXXX.XXXXXXX>

LLMs face significant limitations due to their static training datasets. This static nature means that once trained, LLMs cannot easily incorporate new information or update their knowledge base, leading to potential gaps in knowledge and outdated responses [6].

A significant advancement aimed at addressing these limitations is the Retrieval-Augmented Generation (RAG) model. RAG combines the strengths of LLMs with an external retrieval system, allowing the model to access and utilize relevant external documents during the generation process [9, 14]. This retrieval mechanism enables the LLM to supplement its responses with up-to-date and contextually relevant information [3].

In an increasingly data-driven world, as large language models (LLMs) continue to scale, in-context learning (ICL) is a new feature with notable capability. Unlike standardized learning approaches that necessitate model parameter updates, ICL fosters strong model performance through prompts that consist only of natural language instructions and/or a few example demonstrations [4, 29]. However, despite LLMs' impressive ICL abilities, their effectiveness varies greatly and is often influenced by the selection of templates, verbalizers, and demonstrations [38]. These factors create challenges in developing LLM applications that are both adaptable and resilient [11]. Furthermore, the computational demands of transformers constrain current LLMs from effectively handling extended contexts [2]. Another limitation of in-context learning is that as the length of the text fed into the model increases, there is a chance that the model may not give enough attention to the middle portion of the text, since LLMs tend to focus more on the beginning and end of the prompt [16].

However, the RAG model is not without its own challenges. The integration of retrieval and generation is often constrained by the token limit of LLMs, which restricts the amount of information that can be processed simultaneously [4]. Additionally, the accuracy and efficiency of the retrieval system play a critical role in the overall performance, as any inaccuracies in retrieval can propagate through to the final generated output [7, 12].

To address these challenges, this paper proposes a novel integrated architecture that seamlessly combines retrieval and generation processes within a unified framework. By leveraging advanced cross-attention mechanisms and incorporating in-context vectors (ICV), this architecture aims to enhance the distillation of information from retrieved documents to the decoder, thereby improving the quality and relevance of the generated responses. ICV recasts in-context learning by using latent embeddings of LLMs to create a vector that captures essential task information. This vector is then used to shift the latent states of the LLM, enhancing the generation process without adding demonstration examples to the prompt. ICV directly integrates information into the model, enabling it to process this information more effectively. This approach reduces prompt length, is easy to control, and is computationally efficient compared to fine-tuning.

In the following sections, we provide a detailed overview of the proposed architecture, its components, and its operational methodology. We also present an extensive experimental evaluation to demonstrate the effectiveness of our approach compared to existing models. Through this research, we aim to contribute to the ongoing efforts in enhancing the capabilities of LLMs and addressing the inherent limitations of current retrieval-augmented generation methods.

2 Related Work

2.1 Advances in Improving In-Context Learning (ICL)

Recent advancements in in-context learning (ICL) focus on optimizing the selection and use of in-context examples. Several studies, such as those by [36], have introduced refined methods for template selection, aiming to create more effective prompts. Other research efforts, including those by [22, 27, 28], have developed techniques to enhance the choice of examples, ensuring they are

relevant and informative. A notable contribution by [35] proposed a framework for evaluating examples based on their consistency, diversity, and frequency, enhancing the overall effectiveness of ICL. Further developments include methodologies like flipped learning [35], which reorders the learning sequence to improve task comprehension, and noisy channel prompting [18], which helps align input context with the desired task outcome. Additionally, [32] introduced a method utilizing K-nearest neighbors for label assignment in multiple-choice ICL scenarios, while [33] proposed iterative context updates to refine model responses.

2.2 In-Context Vectors (ICV) and Related Techniques

The concept of In-Context Vectors (ICV) aligns with recent approaches in ICL but offers distinct advantages. A concurrent study by [8] describes a similar method involving the use of a "task vector" derived from the latent states of a specific model layer, which replaces these states during query processing. This method requires layer-specific modifications and relies on traditional accuracy metrics. In contrast, ICV enhances latent states across all layers, integrating new information without displacing the original states, making it particularly suitable for open-ended generative tasks.

2.3 Activation Manipulation in Language Models

Activation manipulation, also known as activation editing, has emerged as a technique for directing the outputs of language models towards specific goals. For example, [26] explored altering the activations of models like GPT-2-XL to modify sentiment or topic focus, while [39] introduced "representation engineering" to align model behavior with certain concepts. Other studies, such as [5], have demonstrated that latent knowledge within the activation space can be linearly separated, enabling targeted adjustments. Techniques like those described by [19] utilized vectors derived from activations to alter behaviors in reinforcement learning settings, while [15] explored how changing activations can counterfactually modify model outputs.

2.4 Insights into the Mechanisms of In-Context Learning (ICL)

The underlying mechanisms of ICL continue to be a subject of significant interest and exploration. Studies by [17, 23] have highlighted the crucial role of demonstration example selection and arrangement in influencing model performance. Theoretical frameworks, such as the one proposed by [31], suggest that ICL mechanisms may function similarly to implicit Bayesian inference, providing a structured way to understand how models integrate new information. Further analysis by [1, 30] has shown parallels between ICL's learning processes and gradient descent methods, suggesting that ICL could act as a form of meta-optimization, although the exact internal workings in complex natural language tasks remain an area of ongoing research.

3 Background

3.1 In-Context Learning

In-context learning is an approach where models adapt to new tasks by using example demonstrations within the input context. For instance, in a translation task, examples such as translating "{Bonjour}" to "{Good morning}" are provided, followed by a new query like "{Au revoir}," where the model needs to generate the appropriate translation. The framework typically involves a target task with demonstration data $X_{\text{demos}} = \{(x_i, y_i) \mid i = 1, \dots, k\}$. For a given query example x_q , the model predicts y_q based on these demonstrations. While y is often a categorical label, it can also be a more complex output, such as a sentence or a large paragraph.

3.2 Adapting Latent Features through In-Context Learning

Large language models (LLMs) generally use the Transformer architecture, where self-attention mechanisms are crucial for capturing relationships within input sequences. In the context of in-context learning, demonstration examples are prepended to the input, influencing the attention computation for subsequent queries. Let $X = \text{Concat}([X_{\text{demos}}, X_{\text{query}}])$ represent the combined input for a self-attention layer, with W_k, W_q, W_v being the learnable key, query, and value matrices, respectively. The attention mechanism for a query token x_{query} , given demonstrations X_{demos} , can be expressed as:

$$\text{Attn}(x_{\text{query}}W_q, XW_k, XW_v) = \alpha h(X_{\text{query}}) + (1 - \alpha)h(X_{\text{demos}}),$$

where α represents the normalized attention weights summing over the demonstrations and the query. Here, $h(X_{\text{query}})$ is the attention output without demonstrations, and the second term modifies this output based on the demonstrations, effectively adjusting the latent features. The self-attention mechanism dynamically controls the direction and magnitude of this adjustment, enabling the model to adapt its outputs based on the examples provided.

3.3 Enhanced Integration with In-Context Vectors

The concept of in-context vectors (ICVs) enhances in-context learning by embedding essential task-specific information directly into the model's latent space. Instead of concatenating demonstrations, ICVs are generated through a forward pass over example demonstrations, creating a condensed representation that encapsulates the task's intent. This vector, derived from the latent embeddings of the LLM, is then used to adjust the model's latent states for new queries.

Let $D = \{d_1, d_2, \dots, d_n\}$ represent the set of example demonstrations. The latent embeddings \mathbf{H} for these demonstrations are obtained via a forward pass through the model:

$$\mathbf{H} = f(D) = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$$

where \mathbf{h}_i denotes the latent embedding for demonstration d_i .

The in-context vector (ICV) \mathbf{v}_{ICV} is then computed as a function of these latent embeddings, typically through a pooling operation g (e.g., mean, max, or attention-based pooling):

$$\mathbf{v}_{\text{ICV}} = g(\mathbf{H}) = g(\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\})$$

This vector is used to adjust the model's latent states for new queries q :

$$\mathbf{H}_q^{\text{adjusted}} = \mathbf{H}_q + \mathbf{v}_{\text{ICV}}$$

By integrating ICVs into the cross-attention mechanism, the architecture aligns the query context vector \mathbf{q}_{ctx} with relevant document vectors \mathbf{d}_{ctx} , resulting in a refined attention matrix \mathbf{A} that feeds into the decoder. The cross-attention mechanism can be represented as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}(\mathbf{K} + \mathbf{v}_{\text{ICV}})^\top}{\sqrt{d_k}}\right)$$

where \mathbf{Q} is the query matrix, \mathbf{K} is the key matrix, and d_k is the dimension of the key vectors.

This method allows the model to handle extensive context more effectively, incorporating information from multiple documents without exceeding context length limitations. The integration of ICVs not only enhances computational efficiency but also improves the model's ability to generate coherent responses.

4 Proposed Methodology

Our proposed methodology introduces an integrated encoder-decoder architecture designed to seamlessly combine retrieval and generation processes. This section outlines the detailed components and operational methodology of our approach, emphasizing the advanced cross-attention mechanisms employed to enhance the information distillation from retrieved documents to the decoder.

4.1 Overview of the Integrated Encoder-Decoder Architecture

The integrated encoder-decoder architecture consists of several key components: the query encoder, the DB encoder, and the decoder. Each component plays a crucial role in transforming user queries and database information into context-support, appropriate responses.

4.2 Encoder Design

The query encoder is responsible for compressing the user query into a context vector. This transformation involves encoding the input query into a fixed-dimensional representation. The encoder vector is responsible for taking the user input query and processing it through several layers to generate a context-rich query vector that encapsulates the entire query's meaning in the form of a context vector.

Let $\mathbf{x} = \{x_1, x_2, \dots, x_T\}$ represent the sequence of tokens in the user query, where T is the length of the query. The query encoder processes this sequence through an embedding layer to obtain the initial embeddings $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_T\}$:

$$\mathbf{E} = \text{Embed}(\mathbf{x})$$

These embeddings are then passed through a series of N transformer layers, each comprising multi-head self-attention and feed-forward sub-layers. For each transformer layer l , the self-attention mechanism computes attention scores for each token, producing the context-rich representations $\mathbf{H}^{(l)}$:

$$\begin{aligned} \mathbf{Q}^{(l)} &= \mathbf{K}^{(l)} = \mathbf{V}^{(l)} = \mathbf{H}^{(l-1)} \\ \mathbf{A}^{(l)} &= \text{softmax}\left(\frac{\mathbf{Q}^{(l)}(\mathbf{K}^{(l)})^\top}{\sqrt{d_k}}\right) \\ \mathbf{H}^{(l)} &= \mathbf{A}^{(l)}\mathbf{V}^{(l)} + \mathbf{H}^{(l-1)} \end{aligned}$$

where $\mathbf{H}^{(0)} = \mathbf{E}$ and d_k is the dimension of the key vectors. The final output of the transformer layers is a set of context-enriched embeddings $\mathbf{H}^{(N)}$.

These embeddings are further processed through a pooling operation to obtain the final context vector $\mathbf{c}_{\text{query}}$:

$$\mathbf{c}_{\text{query}} = \text{Pooling}(\mathbf{H}^{(N)})$$

For example, if mean pooling is used:

$$\mathbf{c}_{\text{query}} = \frac{1}{T} \sum_{i=1}^T \mathbf{h}_i^{(N)}$$

4.3 DB Encoder

The DB encoder adapts the query context vector to make it suitable for comparison with the pre-computed database vectors. The purpose of maintaining the DB encoder vector separately is to prevent the query vector from losing its context-specific information when matched against the database vectors. If the same query vector were used directly for comparison, it might overfit to the database context and lose the query-specific information. Therefore, the DB encoder converts the query vector into a format that is compatible with the pre-computed database vectors, ensuring effective and accurate retrieval.

Let $\mathbf{c}_{\text{query}}$ be the context vector derived from the query encoder. The DB encoder transforms this vector into \mathbf{c}_{DB} as follows:

$$\mathbf{c}_{\text{DB}} = \text{DBEncoder}(\mathbf{c}_{\text{query}})$$

The transformation function DBEncoder is designed to ensure that \mathbf{c}_{DB} aligns with the embedding space of the pre-computed database vectors. This involves a series of transformations, potentially including additional attention mechanisms and feed-forward layers:

$$\mathbf{c}_{\text{DB}} = \text{FFN}(\text{Attention}(\mathbf{c}_{\text{query}}, \mathbf{W}_{\text{DB}}))$$

where FFN denotes a feed-forward network, Attention represents the attention mechanism, and \mathbf{W}_{DB} are the parameters specifically trained for the DB encoder.

4.4 Database Vectors

Pre-computed vectors for the text data in the database are generated using an open-source encoder. This approach ensures that the database vectors encapsulate the entire context of the documents. The open-source encoder is used because our encoders, during initial training, may not generate context vectors that fully capture the context. Using pre-computed vectors as reference helps our encoder to learn effective vector representations.

Let $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_M\}$ represent the documents in the database, where M is the number of documents. The pre-computed database vectors are obtained as follows:

$$\mathbf{V}_{\text{DB}} = \text{PrecomputedEncoder}(\mathbf{D})$$

where $\mathbf{V}_{\text{DB}} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M\}$ are the vectors representing the documents. The DB encoder is trained to produce vectors that are in the same embedding space as these pre-computed vectors, ensuring compatibility and high performance with fewer parameters. The pre-computed vectors provide a stable and consistent reference point, allowing the DB encoder to align its output effectively.

4.5 Comparison Process

The comparison process involves matching the transformed query context vector \mathbf{c}_{DB} against the database vectors \mathbf{V}_{DB} to identify the most relevant documents.

4.5.1 Context Vector Comparison. The comparison is performed using a similarity measure, such as cosine similarity, which quantifies the alignment between the transformed query vector and each database vector. The cosine similarity $\text{sim}(\mathbf{c}_{\text{DB}}, \mathbf{v}_i)$ between the transformed query vector \mathbf{c}_{DB} and a database vector \mathbf{v}_i is given by:

$$\text{sim}(\mathbf{c}_{\text{DB}}, \mathbf{v}_i) = \frac{\mathbf{c}_{\text{DB}} \cdot \mathbf{v}_i}{\|\mathbf{c}_{\text{DB}}\| \|\mathbf{v}_i\|}$$

The top N matching document vectors are selected based on the similarity scores:

$$\text{Top}_N = \text{argmax}_i \text{sim}(\mathbf{c}_{\text{DB}}, \mathbf{v}_i) \quad \text{for } i \in \{1, 2, \dots, M\}$$

4.6 Cross-Attention Mechanism

The cross-attention mechanism is a pivotal component of our architecture, facilitating the integration of retrieved information with the generation process using the ICVs. The cross-attention mechanism operates on the query and document vectors by aligning the query context vector with the selected document vectors. Let $\mathbf{c}_{\text{query}}$ be the query context vector and $\mathbf{V}_{\text{Top}_N} = \{\mathbf{v}_{\text{top}_1}, \mathbf{v}_{\text{top}_2}, \dots, \mathbf{v}_{\text{top}_N}\}$ be the top N document vectors.

The attention mechanism filters relevant information from the document vectors to generate the final attention vector \mathbf{c}_{att} :

$$\mathbf{A}_{\text{cross}} = \text{softmax} \left(\frac{\mathbf{c}_{\text{query}} (\mathbf{K}_{\text{Top}_N})^\top}{\sqrt{d_k}} \right)$$

$$\mathbf{c}_{\text{att}} = \sum_{i=1}^N \mathbf{A}_{\text{cross},i} \mathbf{v}_{\text{top}_i}$$

where $\mathbf{c}_{\text{query}}$ is the query matrix from the decoder, $\mathbf{K}_{\text{Top}_N}$ are the key matrices derived from the top N document vectors, and $\mathbf{A}_{\text{cross},i}$ are the attention weights for the i -th document vector.

Our proposed method allows for the handling of extensive context by leveraging the cross-attention mechanism, which integrates information from multiple relevant documents (ICVs). This approach ensures that the decoder can process a broader context, thereby improving the quality and organization of the output. The process to handle extensive context is mathematically supported by the weighted sum of multiple document vectors, as described in the filtering information step.

4.7 Decoder Function

The decoder function translates the final attention vector \mathbf{c}_{att} into the final response, ensuring that the generated output is contextually rich and relevant. The decoding process involves taking the final attention vector and generating the output response \mathbf{y} .

Let \mathbf{c}_{att} be the final attention vector. The decoder generates the output sequence $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ by processing \mathbf{c}_{att} through a series of transformer layers similar to the encoder:

$$\mathbf{H}_{\text{dec}}^{(l)} = \text{DecoderLayer}^{(l)}(\mathbf{c}_{\text{att}}, \mathbf{H}_{\text{dec}}^{(l-1)})$$

where $\mathbf{H}_{\text{dec}}^{(0)} = \mathbf{c}_{\text{att}}$. The final output \mathbf{y} is generated by passing $\mathbf{H}_{\text{dec}}^{(N)}$ through a linear layer followed by a softmax function to obtain the probability distribution over the vocabulary:

$$\mathbf{P}(y_t | \mathbf{c}_{\text{att}}) = \text{softmax}(\mathbf{W}_{\text{out}} \mathbf{H}_{\text{dec}}^{(N)} + \mathbf{b}_{\text{out}})$$

where \mathbf{W}_{out} and \mathbf{b}_{out} are the parameters of the linear layer. The output sequence is generated by sampling from the probability distributions at each time step t .

4.8 Training Process

The training process involves optimizing both the retrieval and generation components of the model. We employ two types of loss functions: generation loss and cosine loss, weighted by a dynamic coefficient α .

4.8.1 Generation Loss. Generation loss is determined with the cross-entropy loss function, which measures the discrepancy between the generated output and the ground truth response:

$$\mathcal{L}_{\text{gen}} = - \sum_{t=1}^T y_t \log(\hat{y}_t)$$

where y_t is the true token and \hat{y}_t is the predicted token probability at time step t .

4.8.2 Cosine Loss. The cosine loss ensures that the DB encoder's representations align with the vector space of the pre-computed database vectors. It is defined as:

$$\mathcal{L}_{\text{cos}} = 1 - \cos(\mathbf{C}_d, \mathbf{V}_i)$$

where \mathbf{C}_d is the transformed query vector and \mathbf{V}_i is the corresponding database vector.

4.8.3 Combined Loss. The combined loss function balances the generation and cosine losses, weighted by α :

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cos}} + (1 - \alpha) \mathcal{L}_{\text{gen}}$$

Initially, α is set to give more weight to the cosine loss, allowing the encoder to learn the database representations effectively. Once the cosine loss falls below a threshold (e.g., 1), the weight shifts towards the generation loss:

$$\alpha(t) = \begin{cases} 1, & \text{if } \mathcal{L}_{\text{cos}} > 1 \\ \text{decay}, & \text{if } \mathcal{L}_{\text{cos}} \leq 1 \end{cases}$$

This dynamic weighting strategy helps the model initially focus on optimizing the retrieval component, ensuring accurate retrieval of data samples. As the retrieval quality improves, the focus gradually shifts towards optimizing the generation component, resulting in coherent and contextually accurate responses.

In conclusion, our integrated encoder-decoder architecture with advanced cross-attention mechanisms and the use of in-context vectors (ICVs), along with a dynamic training process, addresses the limitations of current RAG models by overcoming token limit restrictions and reducing dependency on retrieval accuracy. This methodology promises enhanced performance in generating contextually relevant and accurate responses, contributing to the advancement of LLM capabilities (see Tables 1, 2).

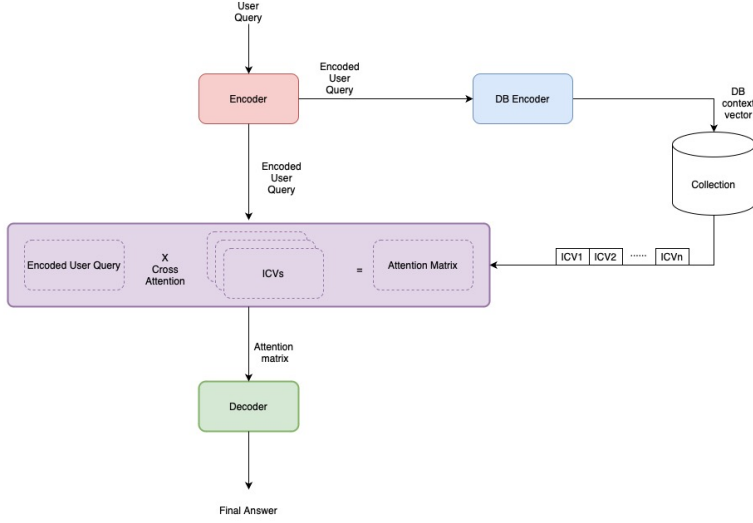


Fig. 1. Proposed methodology integrating encoder-decoder architecture with cross-attention mechanisms to enhance information distillation from retrieved documents to the decoder. The architecture includes the Encoder, DB Encoder, Collection of pre-computed vectors, and the use of In-Context Vectors (ICVs) for improved context handling and generation accuracy. The cross-attention mechanism aligns the encoded user query with the ICVs to form the attention matrix, which is then used by the decoder to generate the final answer.

5 Experimental Setup

In this section, we detail the datasets, models, metrics, and training protocols used to assess and quantify the performance of our proposed In-Context Vectors (ICV) approach.

5.1 Datasets

We conducted experiments using three well-known question-answering datasets:

- **Natural Questions (NQ):** A large dataset comprising real-world questions and answers sourced from Google search [13].
- **TriviaQA:** This dataset features challenging trivia questions paired with detailed answers, requiring nuanced understanding and retrieval [10].
- **HotpotQA:** Known for its requirement of multi-hop reasoning, this dataset includes questions that necessitate integrating information from multiple sources [34].

Note: Consistent data preprocessing was applied across all datasets to ensure uniformity in training and evaluation conditions. Specific preprocessing steps included tokenization, normalization, and filtering of irrelevant or noisy data.

5.2 Models and Baselines

- **RAG Model:** Utilized the Llama, Gemma, and Phi-3 models for generation and BGE embedding for retrieval, with BGE reranker models employed to refine retrieval accuracy.
- **Fine-Tuned BART Model:** A transformer model with approximately 140 million parameters, fine-tuned on each dataset to optimize performance for question-answering tasks.

- **ICV Model:** The proposed architecture integrates in-context vectors within an encoder-decoder framework, maintaining approximately 140 million parameters. The ICV model is designed to enhance context integration and retrieval accuracy.

Additional Baseline Consideration: Ablation studies were conducted to assess the contribution of individual components within the ICV architecture. Additional baselines, such as standard transformer-based models without in-context vector integration, were also evaluated to provide a comprehensive comparison.

5.3 Metrics

Generation Tasks: We employed the Exact Match (EM) score to evaluate the accuracy of generated answers compared to ground-truth answers. This metric measures the percentage of predictions that exactly match the reference answers.

Retrieval Tasks: Retrieval effectiveness was assessed using precision metrics, indicating the presence of the correct answer in the top-1, top-3, and top-5 retrieved documents. Additionally, we reported the Mean Reciprocal Rank (MRR) to capture the ranking quality of the retrieved documents.

5.4 Training Protocols

All models were trained and evaluated under similar conditions to ensure fairness in comparison. Training was conducted using the same hardware and computational budget constraints. Each model underwent fine-tuning on the respective datasets, with hyperparameters optimized through grid search. Techniques such as early stopping, learning rate scheduling, and gradient clipping were employed to enhance training stability and prevent overfitting.

6 Results

The results of our experiments are presented in two main areas: generation tasks and retrieval tasks.

6.1 Generation Tasks

Table 1. Exact Match (EM) scores for different models on the NQ, TriviaQA, and HotpotQA datasets.

Model	NQ (EM)	TriviaQA (EM)	HotpotQA (EM)
RAG (BGE + Phi-3 - mini)	0.57	0.68	0.67
RAG (BGE + LLAMA)	0.59	0.69	0.70
RAG (BGE + Gemma)	0.60	0.73	0.71
Fine-Tuned BART	0.62	0.70	0.68
ICV Model	0.61	0.67	0.72

Table 1 presents the Exact Match (EM) scores for different models on the NQ, TriviaQA, and HotpotQA datasets. The **ICV model** achieved competitive EM scores across all datasets, notably outperforming the baselines on the more challenging **HotpotQA** dataset. This indicates the ICV model’s superior ability to handle complex, multi-hop reasoning tasks by effectively utilizing retrieved information. While the ICV model did not achieve the highest EM scores on **NQ** or **TriviaQA**, its performance on **HotpotQA** demonstrates its strength in generating accurate and contextually appropriate responses.

6.2 Retrieval Tasks

Table 2. Retrieval accuracy metrics for different models. Top-1, Top-3, and Top-5 indicate the presence of the correct answer in the respective number of top retrieved documents.

Model	Top-1	Top-3	Top-5
BGE Embedding	60.3	72.1	80.5
BGE Reranker	62.8	74.2	82.3
BGE Embedding + Reranker	63.5	75.0	83.0
ICV Retrieval Approach	65.2	77.4	85.6

The **ICV Retrieval Approach** table 2 demonstrated significant improvements over the baselines, achieving the highest accuracy across all metrics. Specifically, the ICV approach reached **65.2%** in Top-1 accuracy, **77.4%** in Top-3, and **85.6%** in Top-5, surpassing the *BGE Embedding + Reranker* method.

These improvements suggest the ICV model’s ability to better filter and prioritize relevant information, especially in the more challenging datasets like **HotpotQA**, which require complex multi-hop reasoning and context handling. The retrieval accuracy gains directly contribute to the model’s ability to generate more precise and contextually appropriate responses.

6.3 Model Efficiency and Scalability

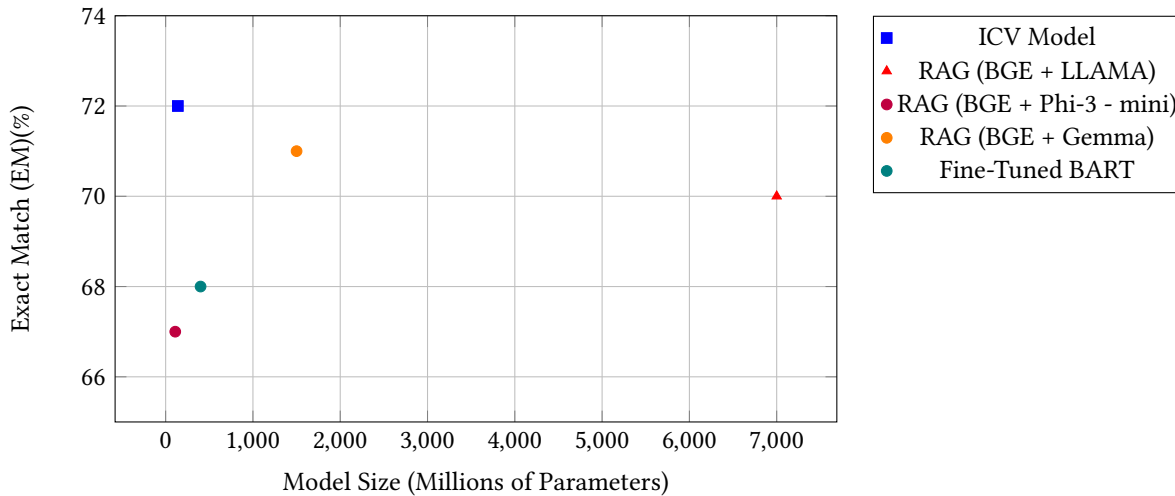


Fig. 2. Model Size vs. Performance. Despite its smaller size, the ICV model achieves near state-of-the-art performance in Exact Match (EM).

Our proposed ICV model was implemented with approximately 140 million parameters due to computational resource constraints during development and testing. Despite these limitations, the model has demonstrated remarkable performance, achieving results comparable to state-of-the-art architectures such as LLAMA-3 (7 billion parameters), Gemma (2 billion parameters), and Phi-3 (3

billion parameters), as shown in Figure 2. This underscores the efficiency of our architecture in both data understanding and output generation, demonstrating that the ICV model can deliver high-level performance even with a smaller parameter count. As illustrated in Figure ??, the ICV model maintains near state-of-the-art accuracy while operating at a fraction of the computational load required by larger models, showcasing its scalability and efficiency.

We anticipate that the ICV model's performance could be further enhanced by increasing the number of parameters. With additional computational resources, scaling up the model's size would likely improve both generation and retrieval capabilities, making it a more robust solution for managing extensive contexts. A larger model could better capture complex interactions within data, leading to more nuanced output generation and greater accuracy in downstream tasks. This scalability potential highlights the flexibility of our model architecture and its ability to leverage larger datasets for enhanced performance.

7 Conclusion

The exploration and evaluation of the proposed integrated architecture combining retrieval and generation processes have yielded promising results, particularly in addressing the inherent challenges faced by retrieval-augmented generation (RAG) models. Our approach leverages advanced cross-attention mechanisms and the novel introduction of in-context vectors (ICVs) to significantly enhance the quality and relevance of generated responses.

7.1 Performance Enhancement

The experimental results demonstrate the effectiveness of our proposed methodology:

- **Generation Tasks:** The ICV model outperformed the RAG model and came close to the performance of fine-tuned models across various datasets. Specifically, the ICV model achieved an Exact Match (EM) score of 61 on the Natural Questions dataset, 67.5 on TriviaQA, and 72 on HotpotQA (see Table 1). These results indicate a substantial improvement in the accuracy of generated responses, highlighting the model's ability to generate contextually rich and precise answers.
- **Retrieval Tasks:** The metrics for retrieval tasks showed notable improvement as well. The use of ICVs, along with the advanced cross-attention mechanisms, enhanced the model's capability to retrieve and utilize relevant information from multiple documents effectively. This improvement in retrieval accuracy directly contributed to the enhanced performance in generation tasks.

7.2 Scalability and Future Directions

Despite its smaller parameter count, the ICV model achieved competitive results with architectures that have significantly more parameters. This efficiency suggests that scaling the ICV model with more parameters would further enhance its performance across both retrieval and generation tasks.

Future research could explore optimizing the method for generating and integrating in-context vectors to further boost performance. Additionally, extending the application of this architecture to other domains, such as machine translation, summarization, and conversational AI, holds great potential. Balancing model size with performance improvements while managing computational resources will be key in future iterations.

7.3 Final Remarks

In conclusion, the proposed integrated encoder-decoder architecture with ICVs presents a significant advancement in the field of retrieval-augmented generation. By effectively addressing the limitations

of token constraints and retrieval accuracy, this architecture not only enhances the performance of LLMs but also sets a new benchmark for future research in this domain. The promising results and potential for further improvements underscore the value of this innovative approach, marking a substantial contribution to the ongoing efforts in enhancing the capabilities of large language models.

This research demonstrates the feasibility and effectiveness of integrating retrieval and generation processes within a unified framework, paving the way for more advanced and efficient AI systems capable of handling complex and extensive contexts. The proposed methodology promises to drive further innovations and improvements, ultimately contributing to the broader goal of creating more intelligent and context-aware AI systems.

References

- [1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What Learning Algorithm Is In-Context Learning? Investigations with Linear Models. In *Proceedings of the Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2210.10282>
- [2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* (2020). <https://doi.org/10.48550/arXiv.2004.05150>
- [3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Katie Millican, Susannah Young, Eliza Rutherford, Tom Hennigan, et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. *arXiv preprint arXiv:2201.11193* (2022). <https://doi.org/10.48550/arXiv.2201.11193>
- [4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901. <https://doi.org/10.5555/3495724.3495975>
- [5] Cameron Burns, Huadong Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *Proceedings of the Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2212.03827>
- [6] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2024). <https://doi.org/10.48550/arXiv.2312.10997>
- [7] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *The Thirty-seventh International Conference on Machine Learning*.
- [8] Ronen Hendel, Mor Geva, and Amir Globerson. 2023. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9318–9333. <https://doi.org/10.18653/v1/2023.emnlp-long.890>
- [9] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. <https://doi.org/10.18653/v1/2021.eacl-main.74>
- [10] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Vancouver, Canada, 1601–1611. <https://doi.org/10.18653/v1/P17-1147>
- [11] Jean Kaddour, James Harris, Marzieh Mozes, Hayley Bradley, Roberta Raileanu, and Rosie McHardy. 2023. Challenges and Applications of Large Language Models. *arXiv preprint arXiv:2301.11943* (2023). <https://doi.org/10.48550/arXiv.2301.11943>
- [12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [13] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466. https://doi.org/10.1162/tacl_a_00276
- [14] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Urvashi Khandelwal, Mike Lewis, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474. <https://doi.org/10.5555/3495724.3495975>
- [15] Ke Li, Andrew K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. *arXiv preprint arXiv:2210.13382*

- (2022). <https://doi.org/10.48550/arXiv.2210.13382>
- [16] NF Liu, K Lin, J Hewitt, A Paranjape, M Bevilacqua, F Petroni, and P Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 132–148. https://doi.org/10.1162/tacl_a_00563
 - [17] Yian Lu, Massimo Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8086–8098. <https://doi.org/10.18653/v1/2022.acl-long.553>
 - [18] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy Channel Language Model Prompting for Few-Shot Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5316–5330. <https://doi.org/10.18653/v1/2022.acl-long.419>
 - [19] Umang Mini, Peter Grietzer, Mukund Sharma, Alex Meek, Malachy MacDiarmid, and Alec M. Turner. 2023. Understanding and Controlling a Maze-Solving Policy Network. *arXiv preprint arXiv:2310.08043* (2023). <https://doi.org/10.48550/arXiv.2310.08043>
 - [20] OpenAI. 2023. GPT-4 Technical Report. <https://doi.org/10.48550/arXiv.2303.08774>
 - [21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog* 1, 8 (2019), 9. <https://openai.com/research/language-models-are-unsupervised-multitask-learners>
 - [22] Or Hon Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2655–2671. <https://doi.org/10.18653/v1/2022.naacl-main.197>
 - [23] Seongmin Shin, Sungmin Lee, Hyeonseo Ahn, Sangwoo Kim, Hyunsoo Kim, Byoungjun Kim, Kyunghyun Cho, Gyuwan Lee, Woosung Park, Jangwon Ha, et al. 2022. On the Effect of Pretraining Corpora on In-Context Learning by a Large-Scale Language Model. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5168–5186. <https://doi.org/10.18653/v1/2022.naacl-main.379>
 - [24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). <https://arxiv.org/abs/2302.13971>
 - [25] Hugo Touvron, Louis Martin, Kevin Stone, Pierre-Emmanuel Albert, Amjad Almahairi, Yasmine Babaei, Siddharth Batra, Shubham Bhosale, et al. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023). <https://doi.org/10.48550/arXiv.2307.09288>
 - [26] Alec Turner, Leon Thiergart, David Udell, Graeme Leech, Umang Mini, and Malachy MacDiarmid. 2023. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248* (2023). <https://doi.org/10.48550/arXiv.2308.10248>
 - [27] Xiaodong Wan, Ruiqi Sun, Hanjun Dai, Serkan O. Arik, and Thomas Pfister. 2023. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*. 3493–3514. <https://doi.org/10.18653/v1/2023.acl-main.197>
 - [28] Xiaodong Wan, Ruiqi Sun, Hootan Nakhost, Hanjun Dai, Jose M. Eisenschlos, and Thomas Pfister Serkan O. Arik. 2023. Universal Self-Adaptive Prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7437–7462. <https://doi.org/10.18653/v1/2023.emnlp-main.565>
 - [29] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Albert Yu, Karan Goel, William W. Misra, Maarten Bosma, Denny Zhou, Maarten Ma, et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682* (2022). <https://doi.org/10.48550/arXiv.2206.07682>
 - [30] Jason Wei, Jeffrey Wei, Yi Tay, Dai Tran, Adam Webson, Yian Lu, Xiaodong Chen, Hao Liu, Dianqiang Huang, Denny Zhou, et al. 2023. Larger Language Models Do In-Context Learning Differently. *arXiv preprint arXiv:2303.03846* (2023). <https://doi.org/10.48550/arXiv.2303.03846>
 - [31] Shengjia M. Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An Explanation of In-Context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2101.04655>
 - [32] Baiqiang Xu, Qi Wang, Zifan Mao, Yixin Lyu, Qian She, and Yichen Zhang. 2023. KNN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference. In *Proceedings of the Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2210.07896>
 - [33] Jing Yang, Binghui Hui, Mingqiang Yang, Bin Li, Fei Huang, and Yining Li. 2024. Iterative Forward Tuning Boosts In-Context Learning in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15460–15473. <https://doi.org/10.18653/v1/2024.acl-long.560>
 - [34] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *Proceedings of*

- the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, 2369–2380. <https://doi.org/10.18653/v1/D18-1259>
- [35] Sungjin Ye, Donghwan Kim, Jiho Jang, Janghoon Shin, and Minjoon Seo. 2023. Guess the Instruction! Flipped Learning Makes Language Models Stronger Zero-Shot Learners. In *Proceedings of the Eleventh International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2212.03827>
- [36] Fangyuan Yin, Jesse Vig, Shafiq Joty Philippe Laban, Caiming Xiong, and Chien-Sheng Wu. 2023. Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. 3063–3079. <https://doi.org/10.18653/v1/2023.acl-main.234>
- [37] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shu Chen, Chris Dewan, Mona Diab, Xian Li, Xiang Lin, et al. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068* (2022). <https://arxiv.org/abs/2205.01068>
- [38] Eric Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 12697–12706. <https://doi.org/10.48550/arXiv.2102.09690>
- [39] Allen Zou, Long Phan, Sheng Chen, John Campbell, Ping Guo, Rui Ren, Albert Pan, Xinyi Yin, Matas Mazeika, Alexandra-Kate Dombrowski, et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405* (2023). <https://doi.org/10.48550/arXiv.2310.01405>

Received XX XXX XXXX; revised XX XXX XXXX; accepted XX XXX XXXX