



A Dynamic Memory Routing Framework for Multimodal Conversational Question Answering in Large Language Models

Walisa Romsaiyud¹

ABSTRACT

Multimodal Question Answering (MQA) in large language models (LLMs) requires adaptive modeling of modality relevance across conversational turns. However, existing approaches rely on static fusion strategies that treat modalities uniformly and fail to capture dynamic modality importance. To address this limitation, we propose DynaRoute, a dynamic memory routing framework for LLM-based MQA. DynaRoute integrates a Bi-LSTM-based conversational memory to model evolving dialogue context and a query-conditioned routing mechanism that dynamically assigns modality relevance at each interaction step. The resulting representations are processed by an LLM-based decoder to generate context-aware responses. Experiments on four benchmarks—VQA-v2, GQA, VisDial, and A-OKVQA—demonstrate consistent improvements over unimodal, static fusion, and mixture-of-experts baselines. DynaRoute achieves an improvement of up to 8.7% under noisy conditions and 6.5% under clean settings, while also obtaining the highest multi-turn consistency (68.9) and robustness (81.3) scores. These results highlight the effectiveness of memory-aware dynamic routing and establish DynaRoute as a principled framework for conversational multimodal question answering.

Article information:

Keywords: Dynamic Memory Routing, Multimodal Question Answering, Conversational AI, Bi-LSTM, Large Language Models

Article history:

Received: December 1, 2025

Revised: March 12, 2026

Accepted: April 22, 2026

Published: April 25, 2026

(Online)

DOI: 10.37936/ecti-cit.2026202.265040

1. INTRODUCTION

Large language models (LLMs) have rapidly become a central paradigm for natural language understanding and generation, demonstrating remarkable capabilities across a wide range of tasks such as text generation, code synthesis, document summarization, and question answering [1–4]. These models rely on dense vector representations learned from large-scale unstructured data, enabling neural networks to capture semantic, syntactic, and contextual patterns with unprecedented accuracy [1,2]. As a result, LLM-based systems have been widely adopted in real-world applications, including intelligent search engines, conversational agents, and decision-support systems in domains such as education, healthcare, and business [8,9]. While early developments in question answering (QA) focused primarily on text-only inputs, recent research has expanded the task into multimodal question answering (MQA), where systems integrate

multiple information sources such as images, textual descriptions, and audio signals [10,12]. Advances in vision–language pretraining and multimodal representation learning have significantly improved performance on standard benchmarks and enabled strong results in object recognition, visual grounding, and cross-modal alignment [11,12]. These developments have established multimodal learning as a critical direction for building more comprehensive and human-like AI systems.

Despite these advances, existing MQA models continue to exhibit notable limitations in conversational and knowledge-intensive scenarios. Most current methods operate on isolated image–question pairs or apply shallow history concatenation when extended to dialogue settings, which restricts their ability to preserve conversational state and track entity references across turns [8,9]. Knowledge-based and embodied question answering have been explored through interactive agents that explore their envi-

¹The author is with the School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi 11120, Thailand, Email: walisa.rom@stou.ac.th

¹Corresponding author: walisa.rom@stou.ac.th

ronments [13]; however, such approaches primarily address navigation and symbolic reasoning rather than conversational multimodal understanding. Although graph-based and hybrid neural frameworks have been proposed to enhance multimodal comprehension [14], they often show limited effectiveness on complex reasoning tasks. Prior work has shown that Bi-LSTM networks improve answer selection by modeling bidirectional semantic dependencies in text-only QA [15,16], yet their effectiveness for representing dialogue-level dependencies in multimodal settings remains underexplored.

Another critical limitation concerns how multimodal information is fused. The majority of current models rely on static fusion strategies, in which modalities are typically combined using fixed or heuristic weights. Although routing mechanisms have recently been introduced in mixture-of-experts (MoE) frameworks [17], such techniques are designed to select computational experts rather than different input modalities. Most existing systems perform routing across subnetworks rather than across heterogeneous modalities such as text, vision, and audio, and typically ignore dialogue history during routing decisions. As a result, modality relevance remains implicit, which prevents models from dynamically re-allocating attention when certain modalities become unreliable due to noise, occlusion, or missing information. This gap highlights a key disconnect between expert routing and adaptive multimodal reasoning.

To address these challenges, we propose DynaRoute, a dynamic memory routing framework for multimodal conversational question answering. DynaRoute explicitly integrates conversational memory with query-conditioned routing to enable adaptive modality selection across dialogue turns. Rather than treating modalities uniformly, the framework learns to prioritize the most informative modality at each interaction step based on both question semantics and dialogue context. This design improves robustness under noisy conditions and enhances interpretability by making modality preferences explicit through routing weights.

The main contributions of this work are summarized as follows:

- (1) We propose a dynamic modality routing mechanism conditioned on question semantics and conversational context, extending routing beyond MoE-style expert selection [17] to modality-level adaptation.
- (2) We integrate a Bi-LSTM-based conversational memory to explicitly model long-range dependencies in multi-turn multimodal QA, building upon sequence modeling capabilities demonstrated in text QA [15,16].
- (3) We design a query-aware routing gate that dynamically prioritizes the most informative modality, addressing limitations of static fu-

sion in multimodal reasoning [10,12,14].

The remainder of this paper is organized as follows. Section 2 reviews related work and formulates the problem setting. Section 3 describes the proposed DynaRoute framework. Section 4 presents experimental results, and Section 5 discusses the findings. Section 6 concludes the paper and outlines future research directions.

2. RELATED WORKS

This section reviews prior work on multimodal question answering, sequential memory modeling with Bi-LSTM, and dynamic routing mechanisms.

2.1 Multimodal Question Answering

Multimodal Question Answering (MQA) aims to answer questions by jointly reasoning across heterogeneous input modalities, such as images, text, audio, and structured knowledge, rather than relying solely on textual information [12,18]. Unlike traditional question answering, where answers are derived solely from documents or databases, MQA requires models to ground linguistic queries in perceptual and contextual evidence from multiple information sources.

An example of multimodal and multi-turn question-answering is shown in Fig. 1.

Q1: Is the man sitting or standing?
 A1: Standing
 Q2: What is the man holding in the image?
 A2: A guitar
 Q3: Is he playing it or just holding it?
 A3: Playing it
 Q4: What is the song's name?
 A4: Heather

Fig. 1: Example of Multimodal and Multi-turn Question-Answering.

Fig. 1 illustrates a multimodal and multi-turn interaction scenario. Visual reasoning is required to infer whether the subject is standing and holding an object, while textual context provides semantic cues about the activity, and audio signals enable identification of the song. Across dialogue turns, the system must track entity states (e.g., holding versus playing) and resolve implicit references, illustrating that MQA requires both cross-modal grounding and temporal reasoning.

MQA often involves additional modalities. In medical applications, systems must integrate radiographic images with clinical reports and laboratory data to infer diagnostic outcomes [18]. In video-based scenarios, models must reason across temporal frames and spoken dialogue to answer questions such as “What did the person do after leaving the room?” or “Why did the character return to the kitchen?” [20]. These examples illustrate that MQA extends beyond

static image understanding and involves complex reasoning across visual, linguistic, temporal, and sometimes knowledge-based cues. Furthermore, when extended to dialogue scenarios, many methods rely on shallow strategies such as history concatenation or fixed-length context windows, which are insufficient for maintaining entity states and resolving references across turns [18, 20]. As a result, models frequently fail to answer follow-up questions such as “*What is it made of?*” or “*Where was it before?*”, because they lack an explicit representation of conversational memory. These limitations reveal two fundamental gaps in existing research: (1) the absence of explicit modeling of conversational context across dialogue turns, and (2) the lack of adaptive modality selection based on question semantics.

This motivates the design of DynaRoute, which integrates conversational memory with query-aware routing to support dynamic and context-sensitive multimodal reasoning.

2.2 Bi-LSTM

Bidirectional Long Short-Term Memory (Bi-LSTM) networks extend recurrent neural networks (RNNs) by modeling sequential data in both forward and backward directions, enabling each token to be conditioned on its full left and right context simultaneously [15–16]. While vanilla RNNs suffer from vanishing gradients and limited long-term dependency modeling, LSTM architectures introduce gated memory cells that regulate information flow through input, forget, and output gates. This gating mechanism allows relevant information to be selectively retained while suppressing irrelevant patterns, thereby improving long-range dependency modeling and training stability in deep sequential architectures [21], as illustrated in Fig. 2.

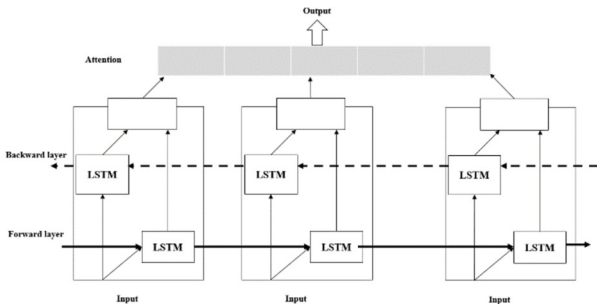


Fig.2: Architecture of a Bidirectional LSTM (Bi-LSTM) with Forward and Backward Recurrent Flows.

Formally, given an input sequence $X = \{x_1, x_2, \dots, x_T\}$, a unidirectional LSTM computes its hidden state at time step t , as shown in Equation (1).

$$h_t = \text{LSTM}(x_t, h_{t-1}) \quad (1)$$

In Bi-LSTM, two independent LSTMs are used.

A forward LSTM processes the sequence from $\{x_1, x_2, \dots, x_T\}$, while a backward LSTM processes the same sequence in reverse. The hidden representation at each time step is obtained by concatenating both directions, as shown in Equation (2).

$$h_t = [h_t^{\rightarrow} || h_t^{\leftarrow}] \quad (2)$$

Equations (1)-(2) show that Bi-LSTM captures both historical and future information, yielding richer representations than unidirectional models. Zhang [21] demonstrated that Bi-LSTM improves response accuracy in educational QA systems by modeling contextual relationships across question sequences. Prabakaran *et al.* [22] further showed that Bi-LSTM captures long-range semantic patterns in automatic script evaluation by modeling both keyword relevance and syntactic continuity.

From a representational learning perspective, Bi-LSTM offers a favorable balance between expressive power and efficiency. Compared with convolution-based sequence models, which focus on local patterns, and graph neural networks, which require explicit relational structure, Bi-LSTM directly learns temporal dependencies from raw sequences [14, 21]. This property makes it particularly suitable for conversational modeling, where dialogue structure is implicit and evolves over time.

This motivates the design of the Bi-LSTM-based conversational memory module in DynaRoute, which extends bidirectional modeling from sentence-level encoding to dialogue-level representation learning.

2.3 Dynamic Memory Routing

Dynamic routing originates from capsule networks, where information is routed between layers based on agreement rather than static weighting [23]. Unlike conventional fusion strategies that merge features, routing mechanisms selectively emphasize representations that are semantically consistent, functioning as a form of structured and learned attention.

Formally, given modality-specific embeddings $h_t^{(k)}$ for each modality $k \in \{\text{image}, \text{text}, \text{audio}\}$, a dynamic routing function computes routing coefficient $\alpha_t^{(k)}$ via a normalized compatibility score, as shown in Equation (3).

$$\alpha_t^{(k)} = \frac{\exp(f(h_t^{(k)}, q_t, M_t))}{\sum_j \exp(f(h_t^{(j)}, q_t, M_t))} \quad (3)$$

Where q_t denotes the encoded question, M_t is the conversational memory at turn t , and $f(\cdot)$ is a learned scoring function implemented by a multilayer perceptron (MLP). This formulation transforms routing into a probabilistic decision process in which each modality is assigned a relevance weight conditioned on the question semantics and dialogue history.

In contrast to traditional dynamic routing protocols studied in network systems, where routing focuses on packet delivery and path optimization [24], routing in neural systems aims to optimize information selection and representation learning. While both paradigms emphasize dynamic adaptability, neural routing operates in semantic space rather than physical networks.

This motivates viewing DynaRoute not merely as a weighting mechanism, but as an adaptive policy that governs evidence selection across modalities. The design enables the model to prioritize visual cues for perception-oriented queries while shifting toward language-based reasoning in knowledge-intensive tasks.

3. THE DYNAROUTE FRAMEWORK

The DynaRoute framework supports dynamic and context-aware multimodal reasoning for conversational question-answering. It consists of six core components: (1) question processing, (2) a Bi-LSTM-based conversational memory, (3) multimodal encoders, (4) a dynamic routing gate, (5) an adaptive multimodal fusion module, and (6) an LLM-based answer decoder, as shown in Fig. 3.

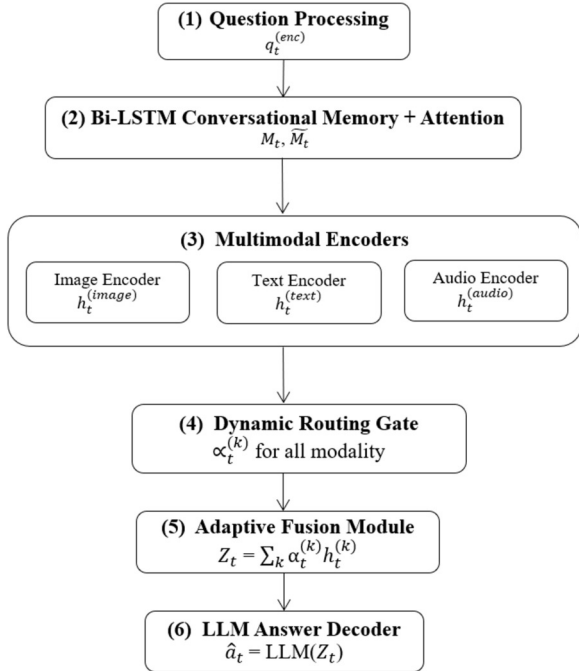


Fig.3: An Overview of the DynaRoute Framework.

3.1 Question Processing

At dialogue turn t , the system receives a natural-language question q_t , which must be transformed into a compact representation that reflects the user’s intent. The processing begins by tokenizing the input question and mapping each token into a continuous

embedding space. These token-level embeddings are then aggregated using a transformer-based encoder, which produces a single semantic vector representing the entire question, as shown in Equation (4).

$$q_t^{(enc)} \in \mathbb{R}^d \quad (4)$$

The vector represents the overall intent of the user’s query, and its semantic cues are implicitly encoded within the dense representation $q_t^{(enc)}$.

For example, given the input question:

$q_t = \text{“What is the man holding in the image?”}$

The ground-truth answer is *“He is holding a guitar.”* At this turn, the system has access to three modalities: (1) an image of a man holding a guitar, (2) a text caption describing the scene, and (3) an ambient audio signal. Because the semantic representation $q_t^{(enc)}$ emphasizes visually grounded elements and action-related semantics, the downstream modules can infer that visual information will be critical for correctly answering this query. Once computed, the encoded question vector $q_t^{(enc)}$ is passed to the Bi-LSTM conversational memory, where it is integrated with information from previous and subsequent dialogue turns.

3.2 Bi-LSTM Conversational Memory

Conversational question answering often spans multiple turns, where the interpretation of the current question depends heavily on earlier user queries and the evolving dialogue context. To model these long-range dependencies, DynaRoute maintains a sequence of encoded questions accumulated over time, as shown in Equation (5).

$$[q_1^{(enc)}, q_2^{(enc)}, \dots, q_t^{(enc)}, \dots] \quad (5)$$

Each element in this sequence corresponds to the semantic embedding of a question asked at a particular dialogue turn. As the conversation progresses, each new question is encoded and appended to this sequence. Rather than processing each question in isolation, the system uses a Bi-LSTM to capture contextual relationships spanning both preceding and subsequent turns.

The forward recurrent pathway models dependencies flowing from earlier turns, as shown in Equation (6).

$$h_t^{\rightarrow} = \text{LSTM}_{\rightarrow} \left(q_t^{(enc)}, h_{t-1}^{\rightarrow} \right) \quad (6)$$

where $q_t^{(enc)}$ is the semantic embedding of the current question and $h_{t-1}^{\rightarrow}$ summarizes the information propagated from all preceding turns.

Conversely, the backward pathway processes the dialogue sequence in reverse temporal order, as shown in Equation (7).

$$h_t^{\leftarrow} = LSTM_{\leftarrow} \left(q_t^{(enc)}, h_{t-1}^{\leftarrow} \right) \quad (7)$$

The hidden state $h_{t-1}^{\leftarrow}$, is derived from the subsequent question during training. Although future turns are not available during inference, the backward LSTM is trained to capture typical follow-up structures and dialogue transitions, enabling it to encode richer contextual patterns. To obtain a unified conversational memory for turn t , the hidden states from the forward and backward pathways are concatenated, as shown in Equation (8).

$$M_t = [h_t^{\rightarrow} || h_t^{\leftarrow}] \quad (8)$$

This memory vector captures both the meaning of the current question and its surrounding conversational context. To emphasize the most relevant dialogue turns, DynaRoute incorporates an attention mechanism over the sequence of Bi-LSTM hidden states. Given the set of concatenated hidden states, as shown in Equation (9).

$$\{h_i = [h_i^{\rightarrow} || h_i^{\leftarrow}]\}_{i=1}^t \quad (9)$$

The attention module computes a relevance weight for each past turn, as shown in Equation (10).

$$\beta_i = \text{softmax}(\mathfrak{v}^T \tanh(W_a h_i)) \quad (10)$$

where W_a and $\mathfrak{v}$ are trainable parameters. These weights indicate how strongly each previous turn contributes to the meaning of the current question. The final attention-enhanced conversational memory is then computed as shown in Equation (11).

$$\tilde{M}_t = \sum_{i=1}^t \beta_i h_i \quad (11)$$

The highlighted formulation of DynaRoute enables the model to selectively focus on earlier dialogue turns that are most semantically aligned with the current question. By explicitly modeling both sequential patterns and attention-based relevance, the Bi-LSTM memory becomes a critical bridge connecting user intent with downstream multimodal reasoning, as described in Fig. 4.

For the running example “*What is the man holding in the image?*”, the conversational memory highlights earlier turns involving the same visual referent, allowing the system to prioritize object-centric modalities in subsequent routing and fusion stages.

3.3 Multimodal Encoders

In the module, the conversational memory M_t and the modality-specific embeddings $h_t^{(k)}$ operate at dialogue turn t , yet they fulfill fundamentally different roles within the DynaRoute framework. While M_t captures the semantic intent of the user’s question along with its multi-turn contextual dependencies, each modality encoder extracts content-specific

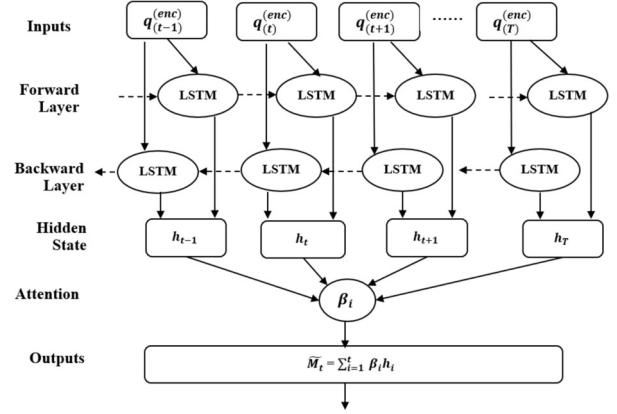


Fig. 4: Bi-LSTM Conversational Memory.

features from its corresponding input stream.

The system receives three heterogeneous inputs at turn t : an image, a text caption, and an audio signal. Each modality is independently transformed into a latent representation suitable for multimodal reasoning, as expressed in Equation (12).

$$h_t^{(k)} \in \mathbb{R}^d, k \in \{image, text, audio\} \quad (12)$$

The resulting vectors $h_t^{(k)}$ constitute a set of modality-specific embeddings. By explicitly separating the conversational memory M_t from these modality embeddings, the framework can determine which modality provides the most relevant evidence for answering the question, rather than treating all modalities as equally informative. This separation improves both the accuracy and robustness of the overall multimodal reasoning pipeline.

For the running example “*What is the man holding in the image?*”, the image encoder identifies high-salience visual objects such as a man, hands, and a guitar, yielding a representation highly aligned with the task. The text caption may contain partial information (e.g., “*A man sitting on a bench*”) and may omit the object of interest, whereas the audio encoder typically provides minimal useful information for visually grounded queries. These heterogeneous embeddings are subsequently passed to the dynamic routing module, where the system selectively determines the relevance of each modality.

3.4 Dynamic Routing Gate

The dynamic routing gate is a key innovation of the DynaRoute framework. Its purpose is to determine how the system selectively prioritizes different modalities when generating an answer. Unlike static fusion strategies, which treat all modalities as equally informative, the routing gate computes context-aware relevance weights based on three complementary signals: (1) the semantic content of the current question, (2) the multi-turn conversational context encoded in

M_t , and (3) the modality-specific feature representations $h_t^{(k)}$ produced at turn t .

At dialogue turn t , each modality encoder produces a latent feature vector $h_t^{(k)}$, where $k \in \{image, text, audio\}$. The routing gate integrates these representations with both the conversational memory M_t and the question embedding $q_t^{(enc)}$. These three signals are concatenated and transformed through a learned linear projection, producing a routing score for each modality, as shown in Equation (13).

$$s_t^{(k)} = W_k \left[h_t^{(k)} || M_t || q_t^{(enc)} \right] \quad (13)$$

where W_k is a trainable projection matrix for modality k . Higher values of $s_t^{(k)}$ indicate strong semantic alignment between the modality and the information required to answer the question. The routing function is implemented as a lightweight multi-layer perceptron (MLP) which parameterizes the projection function W_k in Eq. (13). The MLP maps the concatenated representations into a scalar routing score for each modality, allowing the model to learn nonlinear interactions between question semantics, conversational memory, and modality-specific features.

To enhance context-adaptive decision-making, DynaRoute draws conceptual inspiration from the routing-by-agreement mechanism originally proposed in capsule networks [23]. However, unlike capsule routing, which operates between lower-level and higher-level feature capsules, the proposed routing gate repurposes this idea for multimodal reasoning. Specifically, rather than routing information between neural capsules, DynaRoute dynamically routes across heterogeneous modalities (image, text, audio) based on the current question semantics and multi-turn conversational memory.

The routing scores are normalized via a softmax function, which transforms them into interpretable relevance weights, as shown in Equation (14-15).

$$\alpha_t^{(k)} = \frac{\exp(s_t^{(k)})}{\sum_{j \in \{image, text, audio\}} \exp(s_t^{(j)})} \quad (14)$$

These weights satisfy:

$$\sum_k \alpha_t^{(k)} = 1, 0 \leq \alpha_t^{(k)} \leq 1 \quad (15)$$

Thus, $\alpha_t^{(k)}$ directly represents the relative contribution of modality k to the answer generation process at dialogue turn t .

Although the proposed routing mechanism shares certain similarities with attention-based fusion methods, there are fundamental conceptual differences between the two paradigms. Conventional atten-

tion mechanisms typically compute relevance scores between tokens or feature representations—either within a single modality or across modalities—and aggregate them through weighted summation within a fixed fusion structure. In contrast, the routing gate in DynaRoute performs modality-level decision-making by explicitly determining which modality should contribute more strongly to the final representation.

While attention mechanisms primarily act as feature weighting operations, dynamic routing in DynaRoute functions as an adaptive evidence selection policy conditioned on both conversational memory and question semantics. This design enables the model to dynamically reallocate modality importance across dialogue turns rather than relying on fixed or locally computed attention weights. Consequently, the routing mechanism provides a higher level of control over modality selection compared with conventional attention-based fusion strategies.

This dynamic routing behavior enables the framework to adapt its modality selection per question, per turn, and per conversational context, thereby significantly improving robustness in multimodal reasoning. Fig. 5 illustrates the internal computation flow of the proposed routing gate.

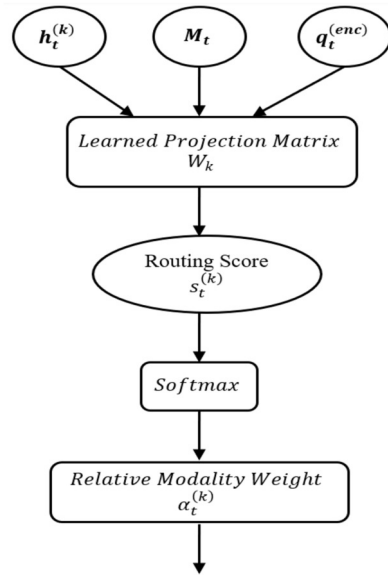


Fig.5: Dynamic Routing Gate.

For the running example, “What is the man holding in the image?” the routing gate may assign the following modality weights:

$$\alpha_t^{(image)} = 0.82, \alpha_t^{(text)} = 0.14, \alpha_t^{(audio)} = 0.04$$

This distribution indicates a strong preference for the image modality. Such a routing decision is consistent with human intuition and demonstrates that the model selectively focuses on the most relevant in-

formation source.

3.5 Adaptive Fusion Module

After the routing gate computes the modality relevance weights $\alpha_t^{(k)}$, DynaRoute constructs a unified multimodal representation by performing a weighted fusion of all modality-specific embeddings. From a theoretical perspective, conventional static fusion strategies assume that all modalities contribute equally or according to fixed heuristic weights. However, in multimodal question-answering tasks, the relevance of each modality often varies depending on the semantic intent of the question and the conversational context. For example, visually grounded questions primarily depend on visual evidence, whereas knowledge-oriented questions may rely more heavily on textual information. Static fusion methods cannot adapt to these variations because the modality weights remain fixed regardless of the input conditions.

In contrast, the proposed adaptive fusion mechanism leverages the routing coefficients $\alpha_t^{(k)}$ to dynamically determine the contribution of each modality. These coefficients are conditioned on the question representation $q_t^{(enc)}$, the conversational memory M_t , and the modality-specific features $h_t^{(k)}$. As a result, the model can suppress noisy or less informative modalities while amplifying the most relevant ones for the current query.

At dialogue turn t , the fused representation Z_t is computed, as shown in Equation (16).

$$Z_t = \sum_k \alpha_t^{(k)} h_t^{(k)} \quad (16)$$

This formulation ensures that modalities with higher routing scores contribute more strongly to the fused embedding. For instance, if the query is visually grounded—such as “*What is the man holding in the image?*”—the visual embedding dominates the fused representation:

$$Z_t = (0.82)h_t^{image} + (0.14)h_t^{text} + (0.04)h_t^{audio} \quad (17)$$

Thus, Z_t serves as a context-aware and question-

driven multimodal representation, integrating only the most relevant evidence while suppressing distractive or irrelevant modalities. This embedding is then passed to the LLM-based decoder to generate the final answer.

3.6 LLM Answer Decoder

The final stage of the DynaRoute framework involves generating the answer from the fused multimodal representation Z_t . DynaRoute employs a large language model (LLM)—such as LLaMA, GPT, T5, or similar architectures—as the answer decoder. The LLM conditions its output on Z_t , treating it as an additional contextual embedding that guides the generation process. The predicted answer at dialogue turn t is produced, as shown in Equation (18).

$$\hat{a}_t = \text{LLM}(Z_t) \quad (18)$$

The decoder generates a natural-language response that reflects both the multimodal evidence and the conversational context captured earlier in the pipeline. For the running example, the system outputs:

$$\alpha_t = \text{“He is holding a guitar.”}$$

This matches the ground-truth answer and demonstrates that DynaRoute successfully (1) identifies relevant modalities, (2) fuses them adaptively, and (3) produces coherent, context-aware responses. The resulting QA pair $(q_t, \tilde{a}_t)$ is appended to the dialogue history and contributes to future turns through the Bi-LSTM conversational memory.

4. EXPERIMENTAL RESULTS AND ANALYSIS

This section evaluates the effectiveness of the proposed DynaRoute framework using four benchmark datasets that collectively cover single-turn question-answering, multi-turn dialogue understanding, and knowledge-intensive reasoning. The experimental protocol is designed to systematically validate the contributions of each core component of DynaRoute.

Table 1: Statistical Summary of Benchmark Datasets Used for DynaRoute Evaluation.

Dataset	Type	#Images	#Dialogues / QA	#Classes	Modality	Description
VQA-v2	Image QA	105,000	505,904	3	Vision + Text	General visual reasoning
GQA	Scene-graph QA	113,018	2M	870 concepts	Vision + Text	Compositional reasoning
VisDial	Visual Dialogue	53,287	1M turns	–	Vision + Dialogue	Multi-turn QA
A-OKVQA	Knowledge QA	13,603	25k	External KB	Vision + Text	Requires world knowledge

4.1 Datasets

In our experiments, we evaluate the proposed DynaRoute framework on four benchmark datasets: VQA-v2, GQA, VisDial, and A-OKVQA [25], which collectively cover visual reasoning, compositional understanding, multi-turn dialogue modeling, and knowledge-based multimodal question-answering, as summarized in Table 1.

4.1.1 VQA-v2

VQA-v2 is a large-scale visual question-answering dataset consisting of 105,000 images and 505,904 question-answer pairs. All questions are categorized into three types: Yes/No, Number, and Other. The dataset focuses on general visual reasoning, requiring models to localize objects, recognize attributes, and reason about spatial relations using both visual and textual information. Due to its diversity and size, VQA-v2 serves as a strong benchmark for evaluating multimodal grounding and answer accuracy.

4.1.2 VisDial

VisDial is a multi-turn visual dialogue dataset containing 53,287 images and approximately 1 million question-answer turns. Each image is associated with a caption and a dialogue sequence between two participants. VisDial is designed to evaluate a model’s ability to maintain dialogue coherence, track references over multiple turns, and integrate visual and linguistic context throughout the conversation. This dataset is therefore particularly relevant for assessing the impact of the proposed Bi-LSTM conversational memory.

4.1.3 A-OKVQA

A-OKVQA is a knowledge-intensive visual question-answering dataset consisting of 13,603 images and approximately 25,000 questions. Unlike conventional QA datasets, many questions in A-OKVQA require external or commonsense knowledge beyond what is visible in the image. This dataset is designed to evaluate a model’s ability to integrate visual perception with knowledge-based reasoning.

4.2 Hardware and Model Configuration

All experiments were conducted under a unified training and evaluation protocol. The proposed DynaRoute framework was implemented using PyTorch 2.1 and the HuggingFace Transformers library (v4.49.0) on Ubuntu 20.04. Training and inference were performed on NVIDIA A100 (40 GB) GPUs with 64 GB of system memory. Mixed-precision training (FP16) was enabled to improve training efficiency and reduce memory consumption.

For the model configuration, BERT-base was adopted as the question encoder, producing semantic representations of dimension $d = 768$. Visual features were extracted using ViT-B/16. Conversational reasoning was modeled using a two-layer Bi-LSTM with a hidden size of 256 for each direction, yielding a conversational memory representation of dimension 512. Attention is applied over hidden states to obtain a context-aware memory vector. The dynamic routing gate was implemented as a three-layer multilayer perceptron that takes as input the concatenation of the question embedding, conversational memory, and modality-specific features.

Training is performed using the AdamW optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized at $2 * 10^{-4}$ with linear warm-up over the first 10% of training steps. Batch size is set to 32, and all models are trained for 15 epochs. A dropout rate of 0.1 is applied to the routing and decoding layers to mitigate overfitting, and gradient clipping is set to 5.0 for training stability.

4.3 The Model Performance Across Benchmark Datasets

This section reports the quantitative performance of the proposed DynaRoute framework across four benchmark datasets representing complementary reasoning capabilities: VQA-v2, GQA, VisDial and A-OKVQA. The complete numerical results are summarized in Table 2, while the relative performance trends are visualized in Fig. 6(a) and Fig. 6(b).

Across all benchmarks, DynaRoute achieves the best overall performance, outperforming unimodal baselines, static fusion strategies, and mixture-of-experts variants. The highest single-dataset score is observed on VQA-v2 (74.7%), reflecting DynaRoute’s

Table 2: Performance Comparison Across Benchmark Datasets.

Model	VQA-v2 Acc	GQA Acc	VisDial NDCG	A-OKVQA Acc	Cons.	Rob.
Text-only	66.2	58.4	54.0	42.1	61.3	72.4
Concat-Fusion	68.5	60.1	54.8	44.7	63.0	74.2
MoE-Uniform	71.4	63.5	56.0	48.0	68.2	78.0
w/o Memory	72.0	63.1	56.6	48.5	68.5	80.0
w/o Routing	72.4	63.0	57.4	49.1	68.4	81.0
DynaRoute	74.7	63.5	57.9	49.1	68.9	81.3

strong capability in visually grounded reasoning tasks where object recognition and semantic alignment are critical. On the knowledge-intensive A-OKVQA dataset, DynaRoute matches the best-performing variant (w/o Routing) at 49.1%, indicating that although routing does not maximize raw accuracy on this dataset, it improves robustness and consistency as shown later in Fig. 6(b). For multi-turn dialogue on VisDial, DynaRoute achieves the highest NDCG score of 57.9, surpassing all baselines including MoE-based models. In contrast, the Text-only baseline consistently yields the lowest performance across all benchmarks, confirming the necessity of multimodal evidence when tackling visually grounded and knowledge-based tasks.

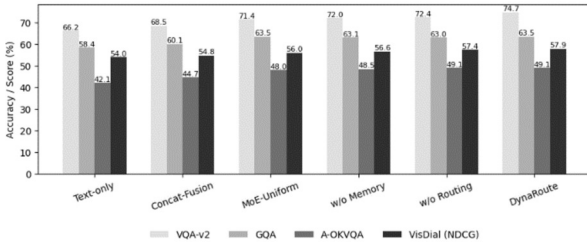


Fig. 6(a): Performance comparison on VQA-v2, GQA, VisDial, and A-OKVQA.

From Fig. 6 (a), DynaRoute consistently outperforms the Text-only and static fusion baselines across all four datasets. The largest performance gain over the Text-only model is observed on A-OKVQA (+7.0%), highlighting the contribution of dynamic routing in knowledge-based reasoning scenarios where visual features alone are insufficient. On GQA, DynaRoute achieves performance comparable to MoE-Uniform (63.5%), indicating that compositional reasoning benefits from expert modeling and structured representations, while still gaining from routing and memory integration. On VisDial, DynaRoute exhibits a clear improvement, demonstrating the effectiveness of conversational memory in maintaining dialogue coherence.

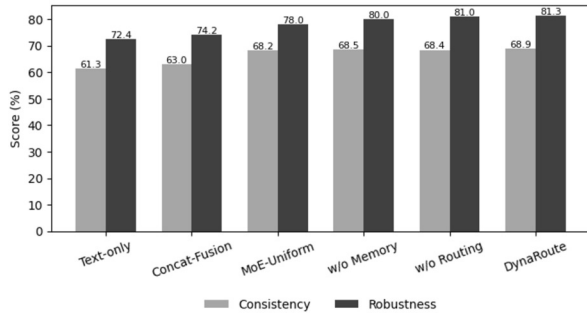


Fig. 6(b): Comparison in Terms of Multi-turn Consistency and Robustness.

From Fig. 6(b), DynaRoute achieves the highest robustness score of 81.3%, whereas the Text-only

model records the lowest score (72.4%), emphasizing the importance of dynamic modality weighting in handling missing or corrupted inputs. In terms of consistency, DynaRoute also attains the highest value (68.9), confirming that Bi-LSTM memory contributes substantially to maintaining dialogue coherence. Although removing memory or routing individually still yields reasonable performance, the full DynaRoute model outperforms all ablated variants, demonstrating that conversational memory and routing mechanisms are complementary rather than redundant.

4.4 Ablation Study

To quantify the contribution of each architectural component, we conduct an ablation analysis by removing individual modules from the full DynaRoute model and evaluating their impact across four benchmark datasets, as illustrated in Fig. 7.

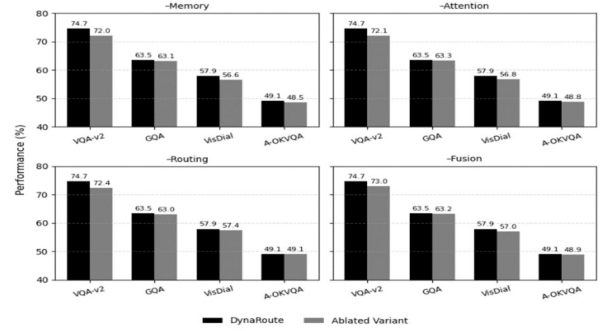


Fig. 7: Ablation Study of DynaRoute Across Four Benchmark Datasets.

From Fig. 7, The ablation results reveal that DynaRoute is a tightly coupled reasoning architecture rather than a collection of independent modules. Removing the Bi-LSTM conversational memory leads to the largest degradation, especially on VisDial, demonstrating that long-range context modeling is critical for entity tracking and discourse continuity in multi-turn conversations. When the attention mechanism is disabled, performance decreases consistently across datasets, indicating that selective focus on dialogue history is necessary to suppress irrelevant context and preserve semantic alignment with the current query. The most significant drops on GQA and A-OKVQA occur when the routing gate is removed, confirming that query-conditioned modality selection is essential for compositional and knowledge-intensive reasoning. Finally, although eliminating the adaptive fusion module results in a smaller decline, it still disrupts performance, suggesting that fusion plays a complementary role by integrating modality-specific representations already shaped by memory and routing. Overall, these results highlight that DynaRoute operates as a coherent reasoning system in which memory, attention, routing, and fusion jointly optimized rather than independently effective.

4.5 Robustness to Noisy and Missing Modalities

We evaluate robustness by comparing DynaRoute with a static fusion baseline under four conditions: clean input, noisy image, masked text, and muted audio, as shown in Fig. 8.

To simulate modality degradation, controlled noise injection procedures were applied during evaluation. For the visual modality, Gaussian noise was added to the image pixels to simulate corrupted visual input. For the textual modality, token masking strategy was applied in which a portion of the caption tokens was randomly replaced with a masking token to represent incomplete textual information. For the audio modality, the audio signal was replaced with silence to simulate missing auditory input.

These perturbations were applied independently to evaluate how the model reallocates modality importance when one of the input sources becomes unreliable. The purpose of this experiment is to examine whether the proposed dynamic routing mechanism can suppress noisy modalities and rely more heavily on the remaining informative ones.

Fig. 8 shows the accuracy comparison between DynaRoute and the static fusion baseline under modality degradation.

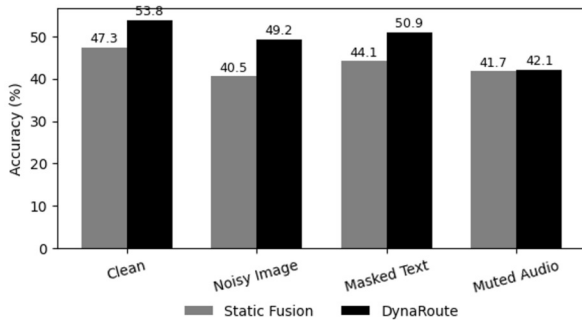


Fig.8: Accuracy Comparison Between DynaRoute and Static Fusion under Modality Degradation.

As shown in Fig. 8, under clean conditions, DynaRoute outperforms static fusion by 6.5 percentage points (53.8% vs. 47.3%). The largest improvement occurs under noisy image conditions, where DynaRoute remains at 49.2% while static fusion drops to 40.5%. Under masked text, DynaRoute maintains 50.9%, outperforming static fusion by 6.8 points. Even under muted audio, DynaRoute retains a slight advantage.

These results demonstrate that DynaRoute maintains reliability when individual modalities degrade. We define static fusion as a baseline that concatenates multimodal representations without learning adaptive weights and consequently lacks the ability to suppress unreliable input, leading to performance degradation under noisy conditions. In contrast, DynaRoute dynamically reallocates modality

importance, resulting in improved robustness.

4.6 Analysis of Dynamic Routing Behavior

To examine whether the proposed routing gate learns meaningful modality preferences, we analyze the average routing weights assigned to each modality across different question types, namely Visual, Textual, and Knowledge-based queries. The distribution of modality weights is illustrated in Fig. 9.

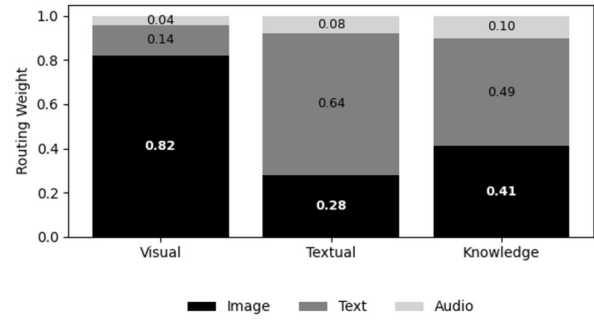


Fig.9: Distribution of Dynamic Routing Weights across Image, Text, and Audio Modalities for Different Question Types.

As shown in Fig. 9, the average routing weights vary across modalities for different question types. For visual queries, image modality weights dominate (0.82). For textual queries, text receives the highest weight (0.64). For knowledge-based queries, weights are balanced between image (0.41) and text (0.49), indicating multimodal dependency. Audio receives consistently low but nonzero weights. These patterns indicate that routing is learned semantically rather than heuristically, confirming that DynaRoute adapts modality selection based on question intent rather than fixed fusion strategies.

Overall, the routing behavior aligns well with human intuition and task-specific requirements. Rather than applying static modality weights, DynaRoute dynamically adapts modality preferences based on the semantic characteristics of each question. These findings empirically validate the effectiveness of the proposed routing gate in learning semantically meaningful and interpretable modality assignment strategies, thereby contributing to more robust and adaptive multimodal reasoning.

4.7 Case Study

The case studies highlight the impact of conversational memory and dynamic routing on robustness and modality selection.

4.7.1 Multi-turn Case Study: Comparison with Static Fusion

This subsection qualitatively compares DynaRoute with a static fusion baseline, focusing on differences

Table 3: Multi-turn Error Analysis Compared with Static Fusion.

Dataset	Turn	Question	Ground Truth	Static Fusion	DynaRoute	Error Analysis
VQA-v2	t ₁	What is the girl holding in her hands?	Red balloon	Nothing	Red balloon	Language bias
	t ₂	Is the balloon bigger than her head?	No	Yes	No	Hallucination
	t ₃	What color is the girl’s dress?	Pink	Red	Pink	Color confusion
GQA	t ₁	What object is the man holding?	Knife	Nothing	Knife	Weak visual grounding
	t ₂	Is the man preparing food?	Yes	No	Yes	Failure on relational reason
VisDial	t ₁	Is the man still eating?	No	Yes	No	No conversational memory
A-OKVQA	t ₂	Why is the man wearing gloves?	Protection	Fashion	Protection	Knowledge not grounded

in entity tracking, modality selection, and error behavior across dialogue turns, as summarized in Table 3.

Table 3 presents multi-turn and knowledge-intensive examples. Static fusion fails due to shallow lexical matching, hallucination, and lack of memory. DynaRoute succeeds by maintaining entity continuity through conversational memory and selecting appropriate modalities through routing. In contrast, DynaRoute consistently produces correct and context-aware responses by jointly exploiting two mechanisms: (i) Bi-LSTM conversational memory, which maintains entity continuity and resolves references across turns, and (ii) dynamic routing, which emphasizes the most relevant modality at each interaction step.

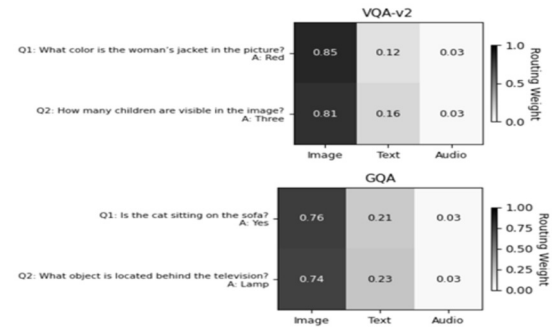
For example, in VQA-v2 and GQA multi-turn interactions, DynaRoute successfully tracks objects introduced in earlier turns and prevents semantic drift, enabling accurate coreference resolution and attribute reasoning. Notably, in knowledge-intensive cases (A-OKVQA), routing weights shift from visual processing to textual and semantic reasoning, allowing the model to perform causal inference, such as recognizing protective intent rather than relying on superficial visual cues.

4.7.2 Routing Behavior Analysis

This subsection analyzes the learned routing weights to determine whether the routing gate assigns modality importance in a semantically meaningful manner. Fig. 10(a)–(b) illustrates the distribution of routing weights across image, text, and audio channels. As shown in Fig. 10(a), the routing weights for VQA-v2 and GQA show that visual queries exhibit strong dominance of the image modality. For instance, in the question “*What color is the woman’s jacket?*”, the image channel receives a weight of 0.85, whereas text and audio obtain only 0.12 and 0.03, respectively. This indicates that the routing gate correctly identifies the task as purely perceptual and ac-

cordingly prioritizes visual evidence. A similar pattern is observed in GQA, where image weights consistently exceed 0.75, confirming that spatial and relational reasoning remains primarily vision-driven despite increased linguistic complexity. This demonstrates that routing behavior is highly structured rather than arbitrary. Visual questions in VQA-v2 and GQA consistently show dominance in the image channel ($\alpha_{\text{img}} > 0.75$).

A distinctly different pattern emerges in VisDial and A-OKVQA, as shown in Fig. 10(b). For example, the question “*Who is well known for playing this instrument?*” shows a stronger text preference (0.55). These results demonstrate that the routing gate correctly shifts focus toward linguistic and knowledge-based reasoning when visual input alone is insufficient.

**Fig.10(a):** Dynamic Routing Weights Across Modalities: VQA-v2 and GQA.

While A-OKVQA cases allocate higher weights to textual reasoning ($\alpha_{\text{text}} > 0.5$), reflecting reliance on world knowledge. In conversational contexts (VisDial), routing remains image-dominant but shows gradual shifts toward text for inferential or temporal queries. These results indicate that routing operates as an implicit attention policy over modalities, learned end-to-end without supervision on modality labels.

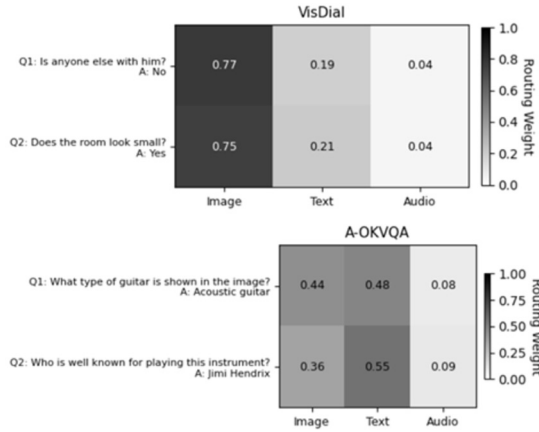


Fig.10(b): Dynamic Routing Weights Across Modalities: VisDial and A-OKVQA.

5. DISCUSSION

5.1 Effectiveness of Dynamic Memory Routing

Experimental results demonstrate that DynaRoute consistently outperforms static fusion and memory-free baselines across all evaluation benchmarks, with particularly strong improvements observed on VisDial and A-OKVQA. This confirms that conversational memory and query-conditioned routing are critical for multi-turn reasoning and knowledge-intensive queries. Unlike conventional multimodal systems that apply fixed fusion strategies, DynaRoute models modality relevance dynamically, enabling adaptive evidence selection based on dialogue context and question semantics.

Furthermore, routing visualizations reveal interpretable modality preferences that align with task demands: visually grounded queries remain image-dominant, while knowledge-oriented questions shift toward textual reasoning. This behavior contrasts with static fusion pipelines, which lack flexibility under noise and ambiguity, and supports the claim that adaptive routing is a principled design for robust multimodal reasoning.

5.2 Limitations and Weaknesses

DynaRoute inherits certain limitations from the underlying pretrained encoders, and performance may degrade when modality representations are noisy or misaligned, particularly in unseen domains. Audio is consistently assigned lower routing weights, reflecting weak dataset-level correlations rather than architectural bias, which suggests the need for richer auditory benchmarks. Although the routing overhead is modest, it may affect latency-sensitive applications. In addition, conversational memory currently operates at the turn level and does not capture token-level interactions, which may limit performance in very long or complex dialogues. These challenges indicate

directions for improving temporal modeling and computational efficiency in future work.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This paper presents DynaRoute, a dynamic memory routing framework for multimodal conversational question-answering in large language models. Unlike static fusion methods, DynaRoute integrates conversational memory, query-conditioned routing, and adaptive fusion to select the most relevant modality at each dialogue turn.

Experimental results on VQA-v2, GQA, VisDial, and A-OKVQA show that DynaRoute consistently outperforms unimodal, static fusion, and MoE baselines. The model achieves strong performance on VQA-v2 (74.7%), VisDial (NDCG 57.9), and knowledge-intensive A-OKVQA (49.1). Moreover, DynaRoute improves robustness under modality degradation by up to 8.7% and yields the highest multi-turn consistency score (68.9), demonstrating its ability to handle noisy input and long-range dependencies. Ablation studies further confirm that memory, routing, and fusion are complementary components. Routing visualizations also show that the model learns interpretable and task-aligned modality preferences. Overall, DynaRoute establishes dynamic routing as an effective and principled paradigm for robust multimodal reasoning in LLMs.

6.2 Future Work

Future work will extend DynaRoute to additional modalities such as video, speech, and sensor data, with particular emphasis on temporal routing for video understanding. Incorporating video streams would enable fine-grained temporal reasoning for activity recognition and long-form video question answering in applications such as surveillance and autonomous systems. Another promising direction is decoder-aware routing, where modality selection is integrated into the LLM to enable token-level control during generation. This would improve factual grounding and reduce hallucination, particularly in long-form responses.

ACKNOWLEDGEMENT

This research was supported by the Institute of Research and Development, Sukhothai Thammathirat Open University, under Grant No. ST-34/2567. The author gratefully acknowledges this support.

AUTHOR CONTRIBUTIONS

W.R. contributed to the conceptualization, methodology, implementation, experimental evaluation, and writing of this study. The author has read and agreed to the published version of the manuscript.

References

- [1] P. Luszczyn and M. Labonne, *LLM Engineer's Handbook*, Packt Publishing, Birmingham, UK, 2024, ch. 1.
- [2] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes and A. S. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.
- [3] Y. Saleh, M. Abu Talib, Q. Nasir and F. Dakalbab, "Evaluating large language models: A systematic review of efficiency, applications, and future directions," *Frontiers in Computer Science*, vol. 7, Art. no. 1523699, 2025.
- [4] V. Alto, *Building LLM Powered Applications*, Packt Publishing, Birmingham, UK, 2024.
- [5] T. Munyer, A. A. Tanvir, A. Das and X. Zhong, "DeepTextMark: A deep learning-driven text watermarking approach for identifying large language model generated text," *IEEE Access*, vol. 12, pp. 40508–40520, 2024.
- [6] Y. Zhou, J. Wen, J. Jia, L. Gao and Z. Zhang, "C-Net: A compression-based lightweight network for machine-generated text detection," *IEEE Signal Processing Letters*, vol. 31, pp. 1269–1273, 2024.
- [7] S. Y.-T. Lee, A. Bahukhandi, D. Liu and K.-L. Ma, "Towards dataset-scale and feature-oriented evaluation of text summarization in large language model prompts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 481–491, Jan. 2025.
- [8] H. Babaei Giglou, G. Rehm, S. Dietze, S. Schimmler and F. Krüger, "Scholarly question answering using large language models in the NFDI4DataScience gateway," in *Natural Scientific Language Processing and Research Knowledge Graphs*, Lecture Notes in Computer Science, vol. 14770, Springer, Cham, Switzerland, pp. 1–18, 2024.
- [9] M. A. Borroto Santana, K. Gallagher, A. Ielo, I. Kareem, F. Ricca and A. Russo, "Question answering with LLMs and learning from answer sets," *Theory and Practice of Logic Programming*, pp. 1–25, 2025.
- [10] Y. Sun, K. Zhang and Y. Su, "Multimodal question answering for unified information extraction," *arXiv preprint arXiv:2310.03017*, 2023.
- [11] T. Qian, R. Cui, J. Chen, P. Peng, X. Guo and Y.-G. Jiang, "Locate before answering: Answer guided question localization for video question answering," *IEEE Transactions on Multimedia*, vol. 26, pp. 4554–4563, 2024.
- [12] M. S. M. Bhuyan, E. Hossain, K. A. Sathi, M. A. Hossain and M. A. A. Dewan, "BVQA: Connecting language and vision through multimodal attention for open-ended question answering," *IEEE Access*, vol. 13, pp. 27570–27586, 2025.
- [13] S. Tan, M. Ge, D. Guo, H. Liu and F. Sun, "Knowledge-based embodied question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11948–11960, Oct. 2023.
- [14] Z. Xiong, L. Zeng, Y. Wu, J. Li, X. Yuan and B. Mo, "Application of deep neural networks integrating multimodal information in intelligent question answering systems," in *Proceedings of the 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, Bristol, United Kingdom, pp. 693–698, 2024.
- [15] M. Bi, Q. Zhang, M. Zuo, Z. Xu and Q. Jin, "Bi-directional long short-term memory model with semantic positional attention for the question answering system," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, Art. no. 77, Sep. 2021.
- [16] A. M. Elema, "Developing Amharic question answering model over unstructured data source using deep learning approach," in *Proceedings of the International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, Bahir Dar, Ethiopia, pp. 108–113, 2022.
- [17] Y. Yuan, A. Gupta, J. Li, S. Dash, F. Wang and M. Zhang, "X-MoE: Enabling scalable training for emerging mixture-of-experts architectures on HPC platforms," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25)*, Association for Computing Machinery, New York, USA, pp. 1315–1331, 2025.
- [18] M. Malinowski, M. Rohrbach and M. Fritz, "Ask Your Neurons: A Deep Learning Approach to Visual Question Answering," *International Journal of Computer Vision*, vol. 125, no. 1–3, pp. 110–135, Dec. 2017.
- [19] H. Demirhan and W. Zadrozny, "Survey of Multimodal Medical Question Answering," *BioMed Informatics*, vol. 4, no. 1, pp. 50–74, 2024.
- [20] Z. Zhao, Z. Zhang, X. Jiang and D. Cai, "Multi-Turn Video Question Answering via Hierarchical Attention Context Reinforced Networks," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3860–3872, Aug. 2019.
- [21] H. Zhang, "Research on Automatic Question Answering System for Online Educational Robots Based on Deep Learning," *Proceedings of the 2024 International Symposium on Intelligent Robotics and Systems (ISOIRS)*, Changsha, China, pp. 279–283, 2024.
- [22] N. Prabakaran, R. Kannadasan, A. Krishnamoorthy and V. Kakani, "A Bidirectional LSTM Approach for Written Script Auto-Evaluation Using Keyword-Based Pattern Matching," *Natural Language Processing Journal*, vol. 5, Art. no. 100033, 2023, ISSN: 2949-

- 7191.
- [23] W. Wang, F. Lee, S. Yang and Q. Chen, "An improved capsule network based on capsule filter routing," *IEEE Access*, vol. 9, pp. 109374–109383, 2021.
 - [24] S. Sirika and S. Mahajan, "Survey on dynamic routing protocols," *International Journal of Engineering Research and Technology*, vol. 5, 2016, Art. no. IJERTV5IS010028.
 - [25] B. S. Kim, J. Kim, D. Lee and B. Jang, "Visual question answering: A survey of methods, datasets, evaluation, and challenges," *ACM Computing Surveys*, vol. 57, no. 10, Art. no. 249, Oct. 2025.



Walisa Romsaiyud is an Associate Professor at the School of Science and Technology at Sukhothai Thammathirat Open University, Thailand. She received her Ph.D. in Information Technology. Her research interests include deep learning, natural language processing, multimodal learning, and question answering systems. Her work focuses on developing intelligent systems that integrate machine learning, conversational AI, and multimodal data for knowledge extraction and decision support. She has been actively involved in research projects related to artificial intelligence applications in education and digital learning environments.