Phase 3: Evaluation

Evaluation Methodology

This phase implemented systematic evaluation of the naive RAG pipeline using four distinct prompting strategies on 100 question-answer pairs from the RAG Mini Wikipedia dataset. For each query, the top 1 retrieved document was passed to the Flan-T5-base language model, and answers were evaluated against ground truth using the HuggingFace SQuAD metric to calculate F1-score and Exact Match.

Prompting Strategies

Four strategies were tested: Basic provides minimal structure with simple context-question-answer formatting. Chain-of-Thought (CoT) encourages step-by-step reasoning with "let's think step by step" prompting. Persona frames the model as a knowledgeable expert assistant. Instruction provides explicit constraints emphasizing context-only information usage with structured guidelines.

Results

Strategy	F1 Score	Exact Match	Avg Time/Query
Basic	53.59%	49.00%	1.11s
Instruction	51.29%	48.00%	1.16s
Persona	51.20%	47.00%	1.07s
CoT	11.73%	0.00%	3.27s

Basic prompting achieved the best performance with F1-score of 53.59 and 49% exact match, outperforming structured approaches by 2.3-2.4 F1 points. Instruction and

persona prompting performed nearly identically around 51 F1. Chain-of-Thought catastrophically failed with only 11.73 F1 and 0% exact match.

Analysis and Hypothesis

The superior performance of basic prompting contradicts intuition about prompt engineering but reveals important insights about Flan-T5's instruction-tuning. The model appears optimized for straightforward question-answering without requiring elaborate prompt scaffolding. Adding explicit instructions, persona framing, or reasoning steps introduces noise that degrades rather than enhances performance on simple factual queries.

Chain-of-Thought's complete failure (0% exact match) indicates fundamental incompatibility between CoT prompting and this task structure. The "let's think step by step" directive triggered verbose reasoning outputs that failed to produce concise factual answers matching ground truth format. For yes/no questions and short factual answers prevalent in this dataset, CoT's elaborative nature works against exact string matching requirements. The processing time of 3.27 seconds per query (versus ~1 second for others) confirms CoT generated significantly longer, more complex outputs that misaligned with evaluation metrics.

The minimal performance gap between instruction and persona prompting (0.09 F1 points) suggests these approaches converge functionally—both add constraint language that slightly hinders the pre-trained model. Interestingly, this finding challenges the assumption that more detailed prompts always improve performance. Because for models like Flan-T5 that underwent extensive instruction-tuning, minimal prompting may suffice.

Sample results reveal persistent challenges even with the best strategy. A query about Abraham Lincoln being the sixteenth President received a retrieval score of 0.71 yet produced an incorrect answer, suggesting either the retrieved passage lacked this specific information or the model failed extraction. However, successful yes/no predictions on questions about the National Banking Act demonstrate the system functions adequately.

The 49% exact match rate indicates roughly half of answers perfectly matched ground truth—a respectable baseline given retrieval limitations. Error analysis suggests two primary failure modes: retrieval failures where top-1 passages lack relevant information (estimated 35-40% of errors), and generation failures where the model produces incorrect answers despite relevant context (estimated 40-45% of errors).