Summarizer Evaluation

The performance of the summarizer is evaluated using the ROUGE (Recall Oriented Understudy for Gisting Evaluation) metric, which includes measures such as ROUGE-1, ROUGE-2, and ROUGE-L for evaluating the quality of the generated summary with a reference summary. ROUGE-1 and ROUGE-2 measures the overlap of unigrams and bigrams respectively between the generated and reference summary, while ROUGE-L measures the overall sequence similarity (Lin, 2004). Firstly, evaluations are performed to find the optimal parameters for summarization. Following this, evaluations are carried out to understand the performance of the summarizer across different cases such as videos of different durations and topics in the domain of residential construction guidance.

Parameter Tuning and Evaluation Metrics

The *chunk size* and *chunk overlap* are the two parameters that were tuned for optimising the summarizer. Both models showed better performance at a chunk size of 2500 and a chunk overlap of 300. This implies the importance of larger chunk size as it includes more information. Table 1 represents the ROUGE scores corresponding to the best parameters.

Model	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Llamini	0.304	0.353	0.327	0.102	0.102	0.102	0.268	0.311	0.288
Llama 2	0.422	0.305	0.361	0.188	0.98	0.129	0.399	0.275	0.325

The summarizer with Llama-2-7b-chat-hf achieves higher recall and f1-score which shows better performance than LaMini-Flan-T5-248M model with the same parameters. This underscores the relevance of model size in the performance. But both the models show low ROUGE-2 scores which indicate lack of bigram overlap.

Evaluation by Video Duration

The performance of the summarizer is evaluated with YouTube videos of different durations. With LaMini-Flan-T5-248M, the F1 scores for ROUGE-1, and ROUGE-L are consistent and around 0.35 for the videos under 30 minutes duration. Although, there is a significant decrease in the ROUGE score for the videos exceeding 30 minutes in duration. The summarizer with Llama-2-7b-chat-hf effectively generates the summary with higher precision and recall. However, there also exists a noticeable drop in the ROUGE score for videos exceeding 30 minutes duration. The summarizer achieved balanced ROUGE scores for shorter videos indicating effective summarization with detail. Figure 1 shows the variation of ROUGE scores for both the models across different video durations.

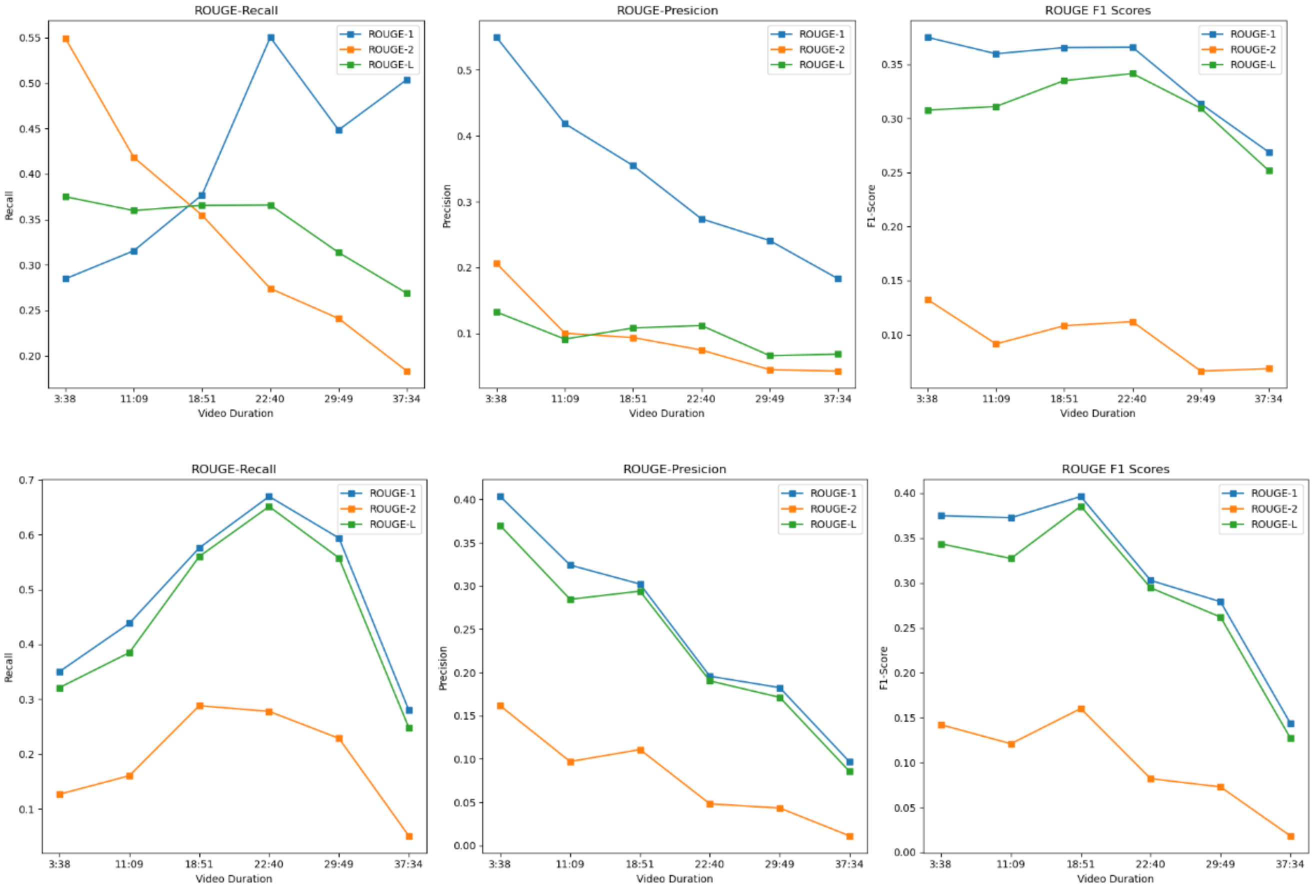

Figure 1. Variations in ROUGE score for summarizer with different video duration, (a) LaMini-Flan-T5-248M and (b) Llama-2-7b-chat-hf

Both the models exhibit similar trend in summarization and shows better performance for shorter videos compared to longer videos. This indicates the challenge in maintaining quality of summarization for longer videos.

Evaluation by Topic

For the evaluation, YouTube videos of various topics in the domain of residential construction guidance such as *"Home-site preparation"*, *"Floor Plan"*, *"Colours and Painting"*, *"Building Permit"*, *"Footings and Foundation"*, *"Framing"*, *"Plumbing"*, *"Building codes and standards"*, *"Types of Building Construction"* were chosen. These topics are significant as they address fundamental aspects of residential construction. Evaluating the summarizer’s performance against YouTube videos of different topics on LaMini-Flan-T5 and Llama-2-7b-chat showed distinct trends. Both the models exhibit lower ROUGE scores for technical topics such as "Framing" and "Types of Building Construction”. It is probably because the complexity of the content making it difficult for effective summarization. LaMini-Flan-T5 particularly showed a low ROUGE score for these topics indicating an apparent drop in the performance. However, with both the models, simpler topics like "Floor Plan" achieved comparably better F1 score. Even though the summarizer performed well across various topics, but the effectiveness of summarization is higher for less technical topics. The quality of summarization is very poor for videos covering complex topics. Figure 2 shows the variation of ROUGE scores for both the models across different video topics.

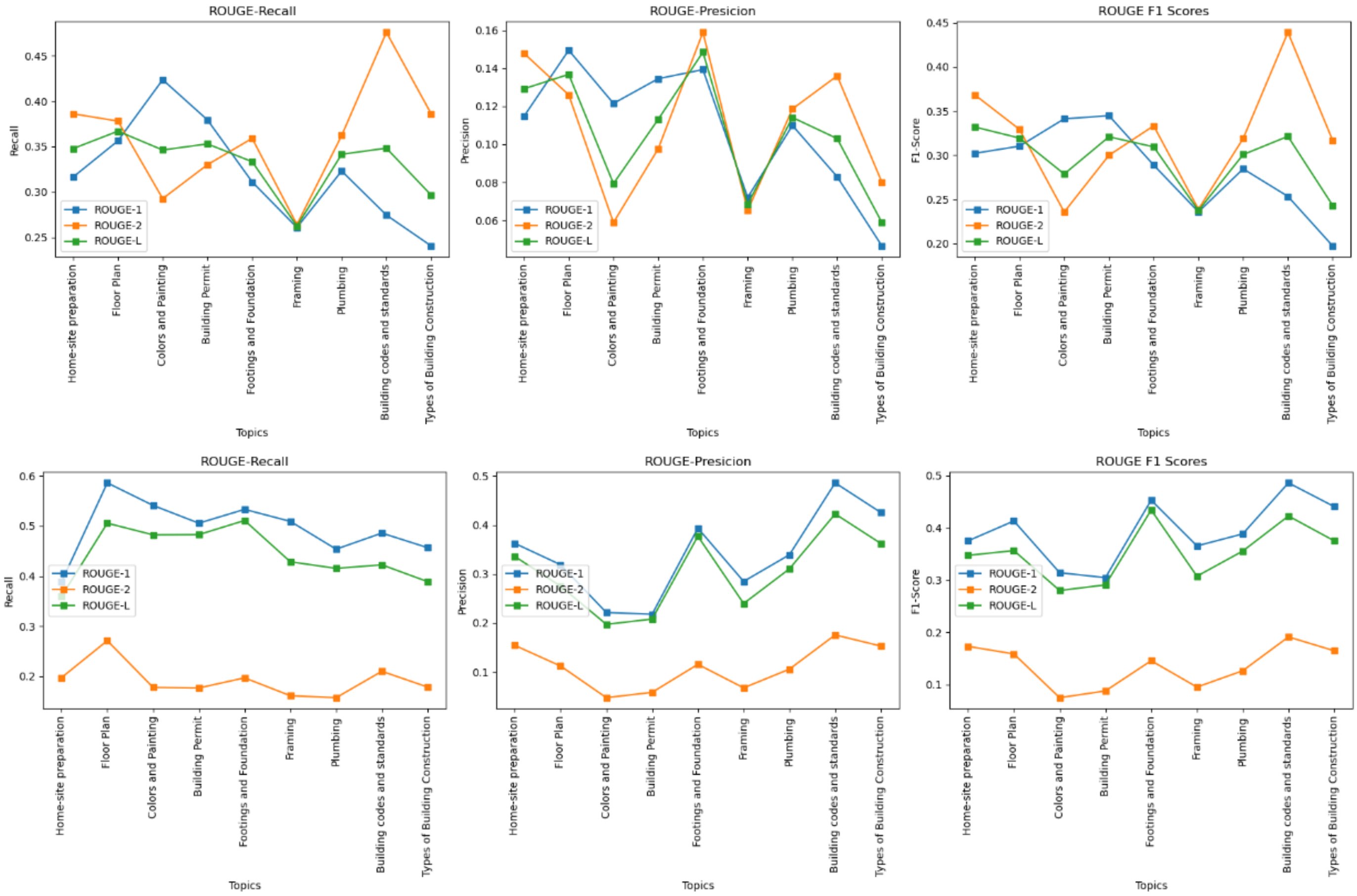

Figure 2Variations in ROUGE score for summarizer with different video topics, (a) LaMini-Flan-T5-248M and (b) Llama-2-7b-chat-hf