Mistral for Text Summarization

Arata Katayama
amk2391

Kian Silva
krs2205

Abstract

This paper evaluates the zero-shot abstractive summarization capabilities of Mistral-7B-Instruct-v0.2 across formal and informal content. We optimized hyperparameter configurations, assessed performance across different text types, and compared Mistral's capabilities with the BART model. Our findings reveal Mistral's superior performance on summarizing formal texts compared to informal content and the importance of tradeoffs between creativity and control through varying decoding strategies. This study aims to build stronger understanding of Mistral's potential for real-world summarization applications by highlighting its strengths and limitations in handling diverse content types without task-specific fine-tuning.

1 Introduction

Text summarization is a natural language processing (NLP) task widely used in various domains, including business, healthcare, education, law, and media, to extract key information from large volumes of content. While the advancement of large language models (LLMs) has revolutionized NLP tasks including text summarization, challenges remain, particularly in ensuring the factual consistency and coherence of generated summaries. This paper evaluates the zero-shot abstractive summarization capabilities of Mistral-7B-Instruct-v0.2 (Mistral V2 7B), a 7-billion-parameter transformer-based model, across diverse text domains.

The problem of text summarization presents several key challenges:

- **Factual consistency**: Many LLMs struggle with hallucination, generating content not present in the original text.
- **Generation Truncation**: Summary generation has a tendency to truncate itself at incomplete points.
- **Evaluation complexity**: Assessment of a generated summary must go beyond syntax-level metrics, incorporating evaluations of semantic and factual accuracy, and coherence to ensure the summary's quality and utility.

Previous work done by Oliveira et al. (Oliveira and Lins, 2024) focused on zero-shot text summarization evaluation of various LLMs, including Mistral V2 7B. Among the six LLMs they tested (including Pegasus, BRIO, and Llama 2 7B, etc.) using zero-shot prompting on the CNN-corpus, Mistral V2 7B outperformed the other models in ROUGE and BERTScores. However, some of the summaries generated by Mistral suffered from hallucination and abrupt cut-off.

Our paper expands on their findings on Mistral V2 7B in the following ways:

- Optimization of hyperparameter configurations, specifically temperature and top-p sampling to explore the impact of various decoding strategies.
- The evaluation of the performance of Mistral V2 7B's summarization across a wide spectrum of textual formality and complexity by testing it on the XSum and Reddit datasets.
- Incorporation of formality classifier to assess stylistic appropriateness of generated summaries.

Our study aims to provide a comprehensive assessment of Mistral V2 7B's zero-shot abstractive summarization capabilities across diverse text domains through hyperparameter optimization and the implementation of a holistic evaluation approach.

To provide a robust benchmark for comparison, we will evaluate the performance of the BART (Bidirectional and Auto-Regressive Transformers)

model on the same datasets. BART has demonstrated state-of-the-art results in various NLP tasks, including summarization. By comparing Mistral V2 7B's performance against BART, we aim to identify the relative strengths and weaknesses of Mistral's zero-shot capabilities compared to an established benchmark.

Through these assessments, our study will offer insights into the model's strengths and limitations in handling different types of content without task-specific fine-tuning, ultimately contributing to our understanding of the model's potential for real-world summarization applications.

2 Methods

2.1 Datasets

We use two distinct datasets to evaluate Mistral V2 7B's zero-shot abstractive summarization performance: XSum and Reddit. The XSum dataset consists of structured, longer-form BBC news articles from 2010 to 2017, providing well-formatted content that is typically formal and factual. In contrast, the Reddit dataset includes posts and comments in a variety of lengths and styles, capturing different forms of spoken English and informal language. Statistics about our datasets are included in Table 1, and example data pairs can be found in the appendix.

Dataset	Total Samples	Average Input Length
XSum	204,045	431.07
Reddit	3,848,330	270

Table 1: Example dataset statistics table.

2.2 Models/Approach

In this study, we evaluate Mistral V2 7B for its performance in text summarization, specifically focusing on their zero-shot capabilities. Zero-shot learning refers to the ability of a model to generate high-quality summaries without having been explicitly fine-tuned on the summarization task. This aspect of model performance is essential for assessing generalization, particularly in real-world scenarios where models often need to handle tasks with limited task-specific training.

Mistral V2 7B utilizes advanced techniques which allow it to process long input sequences with reduced memory usage and faster inference times. These architectural improvements make Mistral V2 7B particularly suited for tasks like text summarization, where the model must effectively handle long contexts while maintaining efficiency. Mistral V2 7B has been fine-tuned on instruction-following tasks, enabling it to generate more task-specific outputs, such as summaries, when prompted. However, for this study, we specifically focus on its zero-shot performance, testing its ability to generate summaries without task-specific training data. The model's strong performance on a variety of benchmarks, including reasoning and reading comprehension, further suggests its ability to condense and summarize textual content effectively in a zero-shot context. (Jiang et al., 2023)

For a benchmark comparison, BART was selected due to its proven effectiveness in sequence-to-sequence tasks, including abstractive and extractive summarization. BART has been shown to perform well in summarization tasks without the need for fine-tuning on specific datasets, making it a reliable baseline for testing zero-shot summarization capabilities. Like Mistral V2 7B, we evaluate BART under its optimal configuration for summarization tasks, ensuring that it is compared to Mistral 7B's best zero-shot performance. This allows us to examine how well both models handle summarization tasks when no task-specific fine-tuning is available, thus providing a direct comparison of their zero-shot abilities. (Lewis et al., 2020)

The goal of this evaluation is to understand the trade-offs between model size, efficiency, and performance in zero-shot summarization. By focusing on the best-performing configuration of Mistral 7B, we hope to better evaluate Mistrals quality of abstractive summaritzation performance in both formal and informal text.

2.3 Experiments (Kian) (20 pts)

The primary goal of our study was to optimize the performance of the Mistral language model for abstractive text summarization. Specifically, we aimed to identify the ideal input text length and prompt structure to produce concise, high-quality summaries while minimizing truncations and hallucinations. We extended the experiments to the BART-Large-CNN model to compare performance across models under similar configurations.

We used the XSum dataset as the primary corpus for our experiments. To address observed truncation issues in Mistral's outputs, we preprocessed the dataset by filtering data points to a maximum length of 1,015 characters. This constraint ensured

that the input text adhered to the model's token limits, improving computational efficiency and the quality of generated summaries. Further along in the study we incorporated the Reddit dataset, applying the same preprocessing steps to maintain consistency across datasets.

After several attempts to identify the best performing prompt for our model, we decided to employ a prompt structure adapted from Oliveira et al. (Oliveira and Lins, 2024). The prompt was formatted as follows:

‘‘‘ Article: [dataset text]
Summarize the article in one sentence.
SUMMARY: ‘‘‘

This prompt effectively reduced hallucinations and overly verbose responses. However, initial evaluations revealed a tendency for summaries to be truncated, highlighting the importance of input length adjustments.

To explore the impact of hyperparameter configurations on the summary quality of Mistral, we varied Top-p (nucleus sampling probability) and Temperature (sampling randomness) as shown in Table 2. Each configuration was tested by generating 100 summaries from both the XSum and Reddit datasets. Given the zero-shot setup, no training or fine-tuning was performed.

We applied a multi-dimensional framework to assess the quality of generated summaries. For lexical similarity, we used ROUGE-1, ROUGE-2, ROUGE-L, and BLEU metrics to measure n-gram overlaps between the generated and reference summaries. For semantic similarity, we employed BERTScore to evaluate the alignment of meaning using Precision, Recall, and F1 scores. Additionally, a formality classifier was utilized to compare the stylistic alignment of generated and reference summaries, offering insights into the tone and structure of Mistral's outputs.

We took Mistrals highest performing hyperparameters to select for further evaluation with the BART-Large-CNN model. The BART model was initialized using the same Hugging Face pipeline and subjected to identical preprocessing and evaluation protocols for direct comparison.

Through this approach, we ensured the reproducibility of our experiments and provided a robust framework for assessing the efficacy of both models in abstractive summarization tasks.

Model Configuration No.	Top-p	Temperature
1	0.8	0.7
2	0.7	0.5
3	0.5	0.5

Table 2: Experiment Hyperparameters

3 Results & Analysis

The following Table 3 and Table 4 shows the text summarization evaluation results for Mistral V2 7B and BART on both XSum and Reddit.

XSum vs. Reddit

Our comprehensive evaluation of Mistral V2 7B's zero-shot abstractive summarization capabilities revealed a clear performance difference between formal and informal text domains. Despite the general advantages of zero-shot learning for model generalization, Mistral V2 7B exhibited significantly stronger performance on formal texts from the XSum dataset, consistently outperforming its results on informal texts across all evaluation metrics. Examples of generated summaries are shown in Appendix A.

Mistral V2 7B achieves significantly higher ROUGE and BLEU scores on formal content. This direct comparison can be seen in Figures 1 and 2. This demonstrates Mistral's ability to generate summaries that retain more relevant information from the reference summaries while maintaining a high degree of lexical similarity. Counterintuitively, despite the substantial differences in the lexical-based metrics, there were clear but small differences seen in BERTScores. This suggests that despite challenges with exact word matching in informal texts, Mistral maintains reasonable semantic understanding across both domains.

The significance in the performance difference between formal and informal text summarization can potentially be associated with one of the architectural characteristics of Mistral V2 7B. Given that it was zero-shot prompted without any fine-tuning, the result implies that Mistral V2 7B was likely trained predominantly on formal structured text. While the training set is not revealed due to proprietary reasons, this inherent bias in pre-training appears to have enhanced the model's proficiency in processing and generating summaries for well-structured, formal content while limiting its effectiveness with informal text.

Model	Config	R1	R2	RL	BERT (P)	BERT (R)	BERT (F1)	BLEU	Formality
Mistral	1	0.130	0.027	0.109	0.827	0.844	0.835	0.0	Formal
Mistral	2	0.146	0.029	0.118	0.836	0.853	0.845	0.0	Formal
Mistral	3	0.101	0.016	0.086	0.843	0.855	0.849	0.0	Formal
BART	1	0.198	0.117	0.169	0.831	0.880	0.854	0.075	Formal

Table 3: Performance Results on Reddit Dataset

Model	Config	R1	R2	RL	BERT (P)	BERT (R)	BERT (F1)	BLEU	Formality
Mistral	1	0.271	0.0822	0.202	0.859	0.879	0.869	0.036	Formal
Mistral	2	0.278	0.0818	0.206	0.862	0.880	0.871	0.038	Formal
Mistral	3	0.298	0.0926	0.214	0.873	0.890	0.881	0.040	Formal
BART	1	0.192	0.0306	0.123	0.838	0.876	0.857	0.012	Formal

Table 4: Performance Results on XSum Dataset

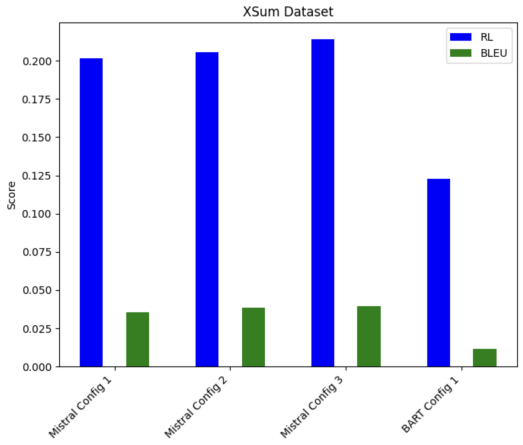

Figure 1: Comparison of RL and BLEU on XSum

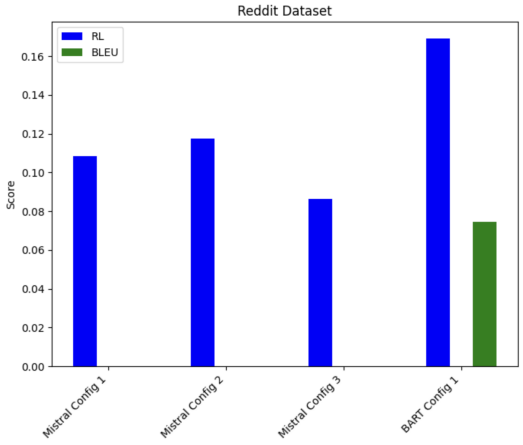

Figure 2: Comparison of RL and BLEU on Reddit

Hyperparameters

Hyperparameter optimization proved crucial in Mistral's performance, with different configurations showing varying effectiveness across datasets. On the XSum dataset, configuration 3 achieved the highest scores across all metrics. For the Reddit dataset, the configuration 2 yielded the highest ROUGE scores, while configuration 3 achieved the highest BERTScore.

For formal content, lower temperature and top-p values consistently produced summaries with better ROUGE, BERTScore, and BLEU evaluations, indicating that stricter decoding parameters produced summaries that better aligned with reference texts both lexically and semantically. This intuitive result suggests that limiting the model's creative freedom through lower temperature and top-p values results in more precise and relevant summarization of formal content.

The analysis of Mistral V2 7B's performance across different hyperparameter configurations on Reddit content reveals a crucial balance between creativity and control in summarizing formal and informal text. Configuration 2 achieved the highest ROUGE scores, suggesting that allowing some flexibility in word choice (higher top-p) while maintaining reasonable constraints (moderate temperature) helps better capture the diverse writing styles and vocabulary typical of informal content.

Conversely, the more conservative configuration yielded the highest BERTScore, indicating that stricter generation parameters are better at preserving semantic meaning and maintaining relevance with the source text. This creates a noteworthy trade-off between lexical matching and semantic

preservation, where moderate creativity improves ROUGE scores, but potentially at the cost of semantic precision.

Comparing with BART

When compared against BART, Mistral V2 7B showed trade-offs between formal and informal texts: it outperformed BART on formal texts on all configurations across all metric scores, but underperformed on informal content on most metrics except for configurations 2 and 3 having a higher BERT precision score than BART. This result reiterates the claim that while Mistral V2 7B excels at handling formal, structured content, it may require additional optimization for processing informal, conversational text.

Formality classification

While the source and reference texts from XSum and Reddit were classified as formal and informal respectively by the formality classifier (Dementieva et al., 2022), Mistral consistently generated formal summaries for both datasets. This reiterates the inherent bias towards formality in Mistral's outputs, likely due to its pretraining on predominantly formal text. This discrepancy highlights challenges in domain generalization and zero-shot performance, emphasizing the need for style-based metrics in evaluation.

4 Related Work (Arata) (15 pts)

(Deutsch et al., 2021) In our study, we primarily focus on evaluating the generated summaries. While ROUGE is useful for measuring textual overlap, it doesn't accurately assess factualilty. Deutsch et al. proposed QAEval, which generates question-answer pairs from a reference summary and evaluates the candidate summary by answering these questions. Unlike traditional metrics like ROUGE and BLEU, QAEval evaluates beyond the lexical similarity, by measuring the overlapping of information, making it more reliable for detecting hallucinations or irrelevant content. This is especially crucial for models like Mistral V2 7B, which have issues with hallucination and abrupt summary endings. This will help us to assess the quality of the model generated summary from a factual aspect.

(Dementieva et al., 2022) Another important evaluation to perform on the generated summaries is their formality. Since our study includes two distinct datasets, one being extracted from a news corpus, and one from social media posts, the structural formality of the summaries must correspond to the input text. Dementieva et al. introduced a formality classifier based on statistical, neural-based, and transformer based approach. Their results showed that the bidirectional LSTM based formality classifier outperformed all other on monolingual experiments, which is what we want to leverage as we are only dealing with English datasets. This evaluation will provide insights into how well Mistral V2 7B handles varying levels of formality and whether it maintains the appropriate tone for each dataset, contributing to a more comprehensive understanding of the model's performance across diverse text domains.

(Oliveira and Lins, 2024) One of the key studies relevant to our work is the research carried out by Oliveira et al. on Mistral V2 7B and other LLMs on the task of text summarization. The study demonstrated strong performance in zero-shot summarization tasks on the CNN corpus which is a structured news dataset. Mistral V2 7B outperformed other directly prompted models, including Llama 2 7B Chat and DaVinci, based on evaluation metrics such as ROUGE-L and BERTScore. Their results highlight Mistral's ability to produce cohesive and concise summaries, which is why we chose Mistral. However, the study also identified areas for improvement, Mistral, like other LLMs on text summarization, have the tendency to abruptly end summary generation as well as including irrelevant and non-factual contents. In our study, we intend to tackle these limitations of the model by exploring the different hyperparameter settings, and if time allows, trying few-shot prompting to enhance the model's summarization performance.

(Lewis et al., 2020) An important part of our evaluation of Mistral's performance relies on the knowledge of other well-documented models such as the BART model. This study builds on the foundations established by BART, a denoising autoencoder for sequence-to-sequence pretraining, which has demonstrated strong performance in both generative and discriminative NLP tasks. BART's flexible pretraining framework, involving text corruption techniques like text infilling and sentence shuffling, has set a benchmark

for model robustness and adaptability across summarization, translation, and comprehension tasks. By integrating a bidirectional encoder and an autoregressive decoder, BART bridges the capabilities of models like BERT and GPT, achieving state-of-the-art results in tasks requiring natural language understanding and generation. Insights gained from BART's design and evaluation inform our approach to Mistral, allowing for a comparative analysis of its efficiency and accuracy in similar contexts.

(Jiang et al., 2023) A core piece of this study is based on the work of Jiangs study on Mistral. Mistral 7B builds upon the advances introduced by these models, including the use of advanced attention mechanisms for efficiency and performance. LLaMA models have demonstrated the utility of open-source pretraining at scale, while BART's sequence-to-sequence architecture highlights the importance of adaptability across diverse NLP tasks. By integrating grouped-query and sliding window attention mechanisms, Mistral 7B refines these principles, achieving state-of-the-art results in tasks like reasoning, mathematics, and code generation, even when compared to larger parameter models. This study further reinforces the importance of efficiency-focused innovations in modern language modeling.

5 Conclusion

This study aimed to evaluate the zero-shot abstractive summarization capabilities of Mistral-7B-Instruct-v0.2 across diverse text domains, focusing on formal and informal content. Our primary objectives were to optimize hyperparameter configurations, holistically assess performance across different text types by incorporating formality classifier, and compare Mistral's capabilities with the BART model. Our findings revealed the following about Mistral V2 7B:

- Demonstrates significantly stronger performance on formal texts compared to informal content
- Despite the large difference in lexical-based metrics, BERTScores showed smaller variations, suggesting that Mistral maintains reasonable semantic understanding across both formal and informal domains.
- Lower temperature and top-p values consistently producing better summaries for formal content
- For informal texts considering the trade-offs between creativity and control affects the summarization performance
- When compared to BART, Mistral V2 7B outperformed on formal texts across all metrics but showed a general under-performance on informal content

Limitations / Future Works

Our study had several limitations. First, the evaluation was limited to English-language datasets, limiting the generalizability of our findings to other languages. Future work could expand this analysis to multilingual summarization tasks. Second, our focus on zero-shot performance, while valuable for assessing generalization, does not explore the potential benefits of fine-tuning Mistral for specific summarization tasks. Lastly, the study was conducted using the free version of Google Colab, which necessitated filtering the datasets to include only shorter pieces of data and limiting us to generating 100 summaries at a time before exhausting the available resources.

6 Contribution Statement

This project was divided and conquered evenly by Kian and Arata. Having split up duties in the code and writing evenly. While Kian worked predominantly on the execution of the Mistral text generation and Arata on the BART text generation along with configuring the evaluation metrics. For this paper, we divided the sections evenly.

References

Daryna Dementieva, Ivan Trifinov, Andrey Likhachev, and Alexander Panchenko. 2022. Detecting text formality: A study of text classification approaches. *arXiv preprint arXiv:2204.08975*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b: A high-performance, efficient language model. *arXiv preprint arXiv:2310.06825v1*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Maziar Mohtashami, Chris Callison-Burch, Jennifer Palomaki, Leandro Tunstall, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Hilário Oliveira and Rafael Dueire Lins. 2024. Assessing abstractive and extractive methods for automatic news summarization. In *Proceedings of the ACM Symposium on Document Engineering 2024*, pages 1–10.

A Appendix

Sample XSum Article (BBC)

The 26-year-old was injured in last week's win over Exeter. Director of rugby Richard Cockerill told BBC Radio Leicester: "With eight weeks of the season to go that is probably the last we will see of him." Meanwhile, scrum-half Sam Harrison, 24, has signed a new deal, but the length of the contract has not been revealed. Benjamin has been plagued by injuries since signing from Worcester in 2012, suffering a serious neck injury which delayed his first-team debut by 15 months. He also sustained a knee injury that ruled him out for four months last term. But Cockerill said: "It is certainly not too serious and he will make a full recovery in the next eight to 10 weeks."

Reference Summary

Benjamin was injured in last week's win over Exeter. Director of rugby Richard Cockerill says he will make a full recovery. Scrum-half Sam Harrison has signed a new deal, but the length of the contract has not been revealed. The 24-year-old has been plagued by injuries since signing from Worcester in 2012.

Mistral's Generated Summary

Leicester Tigers winger Miles Benjamin is likely to be out for the rest of the season because of a knee injury, reports BBC Radio Leicester.

Sample Reddit Post

Art is about the hardest thing to categorize in terms of good and bad. To consider one work or artist as dominate over another comes down to personal opinion. Sure some things maybe blatantly better than other works, but it ultimately lies with the individual. I personally enjoy the work of "street artists" (using quotations not to be sarcastic, but mainly because this is in a different category than graffiti and since my background is not in art I don't know what the "proper" term is , if there is one), but I do see where you are coming from. CLET tends to use the same images continuously (to a point where one could say "Is this it?") as do most street artists (I do think this term is thrown around a lot more than it should be, I agree with you there) and it can be annoying.

Reference Summary

Personal opinions 'n shit.

Mistral's Generated Summary

The article discusses the topic of street art and how it has become increasingly popular in recent years. It also talks about the controversy surrounding certain types of street artists who have been accused of being too commercialized and repetitive.