

Large Malaysian Language Model Based on Mistral for Enhanced Local Language Understanding

Husein Zolkepli*

Aisyah Razak†

Kamarul Adha‡

January 21, 2024

Abstract

In this paper, we present significant advancements in the pretraining of Mistral 7B, a large-scale language model, using a dataset of 32.6 GB, equivalent to 1.1 billion tokens. We explore the impact of extending the context length, releasing models with context lengths of 4096 and 32768 tokens, and further refining performance with a specialized 16384 context length instruction-tuned model, we called it Malaysian Mistral.

Our experiments demonstrate the efficacy of continue pretraining and the influence of extended context lengths on Mistral 7B’s language understanding capabilities. Additionally, we release a model specifically tuned with a 16384 context length instruction, showcasing its potential for capturing nuanced language intricacies.

Furthermore, our research contributes to the benchmarking of Malaysian Mistral against prominent language models, including ChatGPT3.5 and Claude 2.1. We present compelling results indicating Malaysian Mistral’s superior performance on Tatabahasa (Malay grammar) test set, particularly when fine-tuned with instructions.

All models released at [HuggingFace Mesolitica Malaysian Mistral 7B Collection](#).

1 Introduction

The evolution of artificial intelligence (AI) has witnessed transformative breakthroughs, from the introduction of "Attention is All You Need" [1] with the Transformer architecture, to subsequent advancements like GPT-2, and the revolutionary ChatGPT. These models have sparked immense interest and curiosity in the AI landscape, pushing the boundaries of natural language understanding and generation.

In response to this dynamic landscape, Mistral AI emerged, unveiling its initial model, Mistral 7B [2]. Notably, Mistral 7B showcased superior performance, surpassing benchmarks set by Llama 2 13B across various tasks and even outperforming Llama 1 34B on numerous benchmarks. Impressively, it approached the performance of CodeLlama 7B on code-related tasks while maintaining proficiency in English language tasks. However, an identified gap in its capabilities was the limited understanding of Malaysian context.

- **Fine-tuning Mistral 7B:** Utilizing the computational power of 8x A100 GPUs on a Standard_ND96amsr_A100_v4 Azure instance, we conducted extensive fine-tuning on Mistral 7B. The process involved training the model using context lengths of 4096 and 32768 on a substantial 32.6 GB Malaysian context dataset.
- **Multi-turn Instruction-Tuned Model:** Crafting local context multiturn chat dataset using ChatGPT3.5 and ChatGPT4, we employed Meural Machine Translation to translate the dataset. This approach enhances Malaysian Mistral’s proficiency in multi-turn conversations, contributing to its adaptability across a wide range of local context tasks and coding.

*husein@mesolitica.com

†aisyahrazak171@gmail.com

‡kamarul.adha360@gmail.com

2 Related Work

2.1 English-Centric Bias in Large Language Models

The majority of open-source Large Language Models (LLMs) exhibit a significant bias towards the English language, with minimal representation and training on Malay datasets. An analysis of the widely utilized Common Crawl dataset reveals a mere 0.0742% contribution from the Malay language based on [CC-MAIN-2023-50](#) index. This English-centric bias poses a substantial challenge for applications requiring robust language understanding in Malay, prompting the need for dedicated research and development in this domain.

2.2 Existing Malay Language Models

While the Malay natural language processing (NLP) landscape lacks a dedicated Large Language Model, notable efforts have been made by Mesolitica in the development of specific Malay language models. Notable among these are the Malay Causal Language Model, Malay T5, and Malay Masked Language Model. These models, while contributing significantly to the Malay NLP toolkit [3], are distinct from comprehensive Large Language Models and have limitations in capturing extensive context and nuances.

2.3 Absence of a Malay Large Language Model

Despite the existence of specialized models for Malay, a notable gap remains in the absence of a dedicated Malay Large Language Model. The current state of affairs hinders the progress of research and applications requiring a deeper understanding of the Malay language. A comprehensive Large Language Model for Malay is essential to bridge this gap, enabling advancements in various natural language processing tasks and fostering the inclusive representation of Malay in the AI landscape.

3 Pre-Training Procedure

3.1 Public Data

3.1.1 Wikipedia

We initiated the process by downloading the Malay Wikipedia dump from <https://dumps.wikimedia.org/>. This dump serves as a valuable resource for capturing diverse linguistic contexts in the Malay language.

The pertinent information from the Malay Wikipedia dump was extracted using the <https://github.com/attardi/wikiextractor> tool. This tool will parse XML files into JSON line files and removed unnecessary XML tags.

We also obtained the English Wikipedia dataset from <https://huggingface.co/datasets/wikipedia>. Subsequently, we filtered documents containing keywords 'malay' or 'malaysia'. This targeted filtering ensures that the English dataset retains content relevant for the Malaysian context.

3.1.2 Malay Language study articles

To enrich our dataset, we incorporated the Malay dictionary, specifically the "Kamus Dewan Edisi Keempat." This authoritative source provides a comprehensive collection of Malay words, enhancing the language model's understanding of vocabulary and linguistic nuances.

We also included articles from [JendelaDBP](#). The selected articles cover various domains, including language (bahasa), literature (sastera), society (masyarakat), culture (budaya), economy (ekonomi), and Islamic studies (islam). This inclusion ensures that the language model is exposed to a broad spectrum of content, promoting a more holistic understanding of the Malay language within different contexts.

3.1.3 Malaysia Government public documents

For a comprehensive understanding of legislative proceedings, we incorporated data from the Malaysia Hansard. This authoritative source provides official transcripts of parliamentary debates, contributing to the language model’s exposure to formal and legal discourse.

Legal documents from <https://lom.agc.gov.my> were included to provide the language model with insights into legal terminology, regulations, and official language usage within the Malaysian legal context.

To enhance the language model’s grasp of political discourse, we utilized parliamentary records from <https://parlimen.gov.my>. This inclusion ensures exposure to discussions on national policies, legislation, and political matters.

To incorporate real-world data and statistics, we accessed datasets from <https://data.gov.my>. This enriches the language model’s training data with diverse information on various aspects of Malaysian governance and public services.

Articles from the Mufti Wilayah contribute to the language model’s understanding of religious discourse, ethical discussions, and Islamic perspectives on various topics.

To capture a broad range of government-related documents, we performed a systematic Google search for public PDFs from the ‘gov.my’ and ‘edu.my’ domains. This ensures inclusion of publicly available official documents, reports, and publications.

3.1.4 Malaysia online articles

To augment our dataset with a diverse range of Malaysian context, we employed web scraping techniques to gather public articles covering various topics. This comprehensive approach includes content related to gaming, education, blogging, politics, economy, sports, and a multitude of other subjects. The goal is to provide the language model with a broad and nuanced understanding of Malaysian perspectives, opinions, and discussions across a spectrum of domains. Complete list at [link here](#).

3.2 Deduplicating Data

To minimize redundancy within our dataset, we implemented the MinHash algorithm utilizing the implementation available at <https://github.com/ChenghaoMou/text-dedup>.

We configured the MinHash algorithm with the following parameters:

Parameter	Value
num_perm	256
threshold	0.95
hash_func	sha1
hash_bits	64

Complete deduplicating data implementation at [here](#). All deduped dataset published at [malaysia-ai/dedup-text-dataset](#).

3.3 Postprocessing Data

After the initial data collection and deduplication steps, we applied postprocessing techniques to refine the dataset for optimal training. The following steps were implemented:

- Removal of Texts with HTTP Errors.
- Filtering Texts by Length, we removed texts less than 3 characters.
- Normalization of Whitespace, we replaced 6 spaces or more with 6 spaces.
- Normalization of Punctuation, we replaced 6 dots or more with 6 dots.

Complete postprocessing data implementation at [here](#).

3.4 Pre-Training phase

3.4.1 Tokenizing Data

We adopted a packing technique for tokenization, wherein the dataset was segmented into sequences with context lengths of 4096 and 32768. The sequences were separated by the End-of-Sequence (EOS) token, indicating the end of one segment and the beginning of the next.

Complete tokenizing data implementation at [here](#).

3.4.2 4096 context length

We utilized Standard_ND96asr.v4 Azure instance which contains 8x A100 80 GB GPUs (NVLink 3.0). The pre-training hyperparameters are detailed below:

Hyperparameter	Value
DeepSpeed	ZeRO-3 Offload
Batch Size	20
Learning Rate	constant 2e-5
Precision	bfloat16

Complete pre-training 4096 context length implementation at [here](#).

3.4.3 32768 context length

For 32768 context length, we only trained on 10% of the total dataset and use latest checkpoint from 4096 context length pretrained model from section 3.4.2. We utilized Standard_ND96asr.v4 Azure instance which contains 8x A100 80 GB GPUs (NVLink 3.0). The pre-training hyperparameters are detailed below:

Hyperparameter	Value
DeepSpeed	ZeRO-3 Offload
Batch Size	3
Learning Rate	constant 2e-5
Precision	bfloat16

Complete pre-training 32768 context length implementation at [here](#).

4 Supervised Instruction Fine-tuning Procedure

4.1 Generating Instruction Dataset

In our synthetic instruction dataset creation process, we employed ChatGPT3.5 and ChatGPT4 to generate diverse sets of instructions tailored to specific domains. ChatGPT3.5 was utilized to generate instructions with a focus on Malaysian context, contributing to a dataset enriched with localized language nuances. Additionally, ChatGPT3.5 was employed to generate pseudo Retrieval-Augmented Generation Multi-Turn instructions on Malaysian context, further diversifying the content pool.

For coding instructions and synthetic CommonQA, we leveraged the capabilities of ChatGPT4, utilizing its advanced language generation capabilities to create instructions that specifically pertain to coding tasks. This inclusion broadened the dataset’s scope to encompass both language-related and technical instructions.

A notable aspect of this generation process was the tendency of these models to produce instructions in Indonesian. To address this, we implemented neural machine translation, a powerful tool that seamlessly translated the instructions from Indonesian to standard Malay. This translation system demonstrated proficiency not only in translating between Indonesian and standard Malay but also in handling other

languages such as English, Javanese, and Banjarese. Notably, it exhibited the ability to maintain the structural integrity of the original text, selectively translating only necessary components, such as programming code snippets, while preserving the overall text structure. We open-sourced the neural machine translation model at [mesolitica/translation-t5-small-standard-bahasa-cased-code](https://github.com/mesolitica/translation-t5-small-standard-bahasa-cased-code).

4.1.1 OpenQA MS Wikipedia

We use data from 3.1.1 and generate synthetic OpenQA dataset by using prompt below,

```
1 # l['paragraph'] is wikipedia paragraph
2 s = f"""
3 paragraph ''
4 {l['paragraph']}
5 ''
6
7 berdasarkan paragraph, jana soalan melayu dan jawapan melayu
8 """
9 # pass 's' to ChatGPT3.5 API
```

We use ChatGPT3.5 API to generate the synthetic data, an example of generated dataset,

```
1 {'paragraph': 'The Legend of Korra ialah sebuah siri animasi televisyen Amerika Syarikat yang
2 telah ditayangkan di ...',
3 'url': 'https://ms.wikipedia.org/wiki?curid=823980',
4 'qa': {'qa': [{'question': 'Apakah siri animasi yang ditayangkan di Nickelodeon sejak 2012?',
5 'answer': 'The Legend of Korra'},
6 {'question': 'Siapakah pencipta siri animasi The Legend of Korra?',
7 'answer': 'Bryan Konietzko dan Michael Dante DiMartino'},
8 {'question': 'Apakah yang dimaksudkan dengan 'bending' dalam siri animasi The Legend of
9 Korra?',
10 'answer': 'Kekuatan untuk memanipulasi elemen seperti air, bumi, api, atau udara'},
11 {'question': 'Siapakah Avatar Korra?',
12 'answer': 'Pengganti Aang dalam siri sebelumnya yang menghadapi pergolakan politik dan roh
semangat dalam dunia pemodenan'},
{'question': 'Apakah kejayaan siri The Legend of Korra?',
'answer': 'Kejayaan yang kritikal dan komersial dengan jumlah penonton tertinggi bagi siri
animasi di Amerika Syarikat pada tahun 2012'}]}}
```

We also replicated the same technique for Malaysia Hansard and MS CommonCrawl samples. All synthetic dataset and implementation published at [mesolitica/chatgpt-malaysian-open-qa](https://github.com/mesolitica/chatgpt-malaysian-open-qa).

4.1.2 Malay Instruction with Malaysian context

We follow evolution instruction from [4] with slightly changes,
Original breadth instruction,

```
1 base_instruction = "I want you act as a Prompt Creator.\r\n\  
2 Your goal is to draw inspiration from the #Given Prompt# to create a brand new prompt.\r\n\  
3 This new prompt should belong to the same domain as the #Given Prompt# but be even more  
   rare.\r\n\  
4 The LENGTH and complexity of the #Created Prompt# should be similar to that of the #Given  
   Prompt#.\r\n\  
5 The #Created Prompt# must be reasonable and must be understood and responded by humans.\r\n\  
6 '#Given Prompt#', '#Created Prompt#', 'given prompt' and 'created prompt' are not allowed to  
   appear in #Created Prompt#\r\n"
```

Our breadth instruction,

```
1 base_instruction = "I want you act as a Malay Prompt Creator.\r\n\  
2 Your goal is to draw inspiration from the #Given Prompt# to create a brand new prompt in  
   malay language and malaysia related if possible.\r\n\  
3 This new prompt should belong to the same domain as the #Given Prompt# but be even more  
   rare.\r\n\  
4 The LENGTH and complexity of the #Created Prompt# should be similar to that of the #Given  
   Prompt#.\r\n\  
5 The #Created Prompt# must be reasonable and must be understood and responded by humans.\r\n\  
6 '#Given Prompt#', '#Created Prompt#', 'given prompt' and 'created prompt' are not allowed to  
   appear in #Created Prompt#\r\n"
```

Original depth instruction,

```
1 base_instruction = "I want you act as a Prompt Rewriter.\r\n \  
2 Your objective is to rewrite a given prompt into a more complex version to make those famous  
   AI systems (e.g., chatgpt and GPT4) a bit harder to handle.\r\n \  
3 But the rewritten prompt must be reasonable and must be understood and responded by  
   humans.\r\n \  
4 Your rewriting cannot omit the non-text parts such as the table and code in #The Given  
   Prompt#: . Also, please do not omit the input in #The Given Prompt#. \r\n \  
5 You SHOULD complicate the given prompt using the following method: \r\n\  
6 {} \r\n\  
7 You should try your best not to make the #Rewritten Prompt# become verbose, #Rewritten  
   Prompt# can only add 10 to 20 words into #The Given Prompt#. \r\n\  
8 '#The Given Prompt#', '#Rewritten Prompt#', 'given prompt' and 'rewritten prompt' are not  
   allowed to appear in #Rewritten Prompt#\r\n"
```

Our depth instruction,

```
1 base_instruction = "I want you act as a Malay Prompt Rewriter.\r\n \
2 Your objective is to rewrite a given prompt into malay language, a more complex version and
   malaysia related if possible to make those famous AI systems (e.g., chatgpt and GPT4) a
   bit harder to handle.\r\n \
3 But the rewritten prompt must be reasonable and must be understood and responded by
   humans.\r\n \
4 Your rewriting cannot omit the non-text parts such as the table and code in #The Given
   Prompt#: . Also, please do not omit the input in #The Given Prompt#. \r\n \
5 You SHOULD complicate the given prompt using the following method: \r\n\
6 {} \r\n\
7 You should try your best not to make the #Rewritten Prompt# become verbose, #Rewritten
   Prompt# can only add 10 to 20 words into #The Given Prompt#. \r\n\
8 '#The Given Prompt#', '#Rewritten Prompt#', 'given prompt' and 'rewritten prompt' are
   not allowed to appear in #Rewritten Prompt#\r\n"
```

We use ChatGPT3.5 API to generate the synthetic data, an example of generated dataset,

```
1 Instruction: Tulislah sebuah cerita pendek dengan gaya narasi orang ketiga tentang seorang
   protagonis yang harus membuat keputusan karier penting. Ceritakan bagaimana protagonis
   tersebut menghadapi dilema antara mempertahankan pekerjaan yang stabil atau mengambil
   risiko dengan mencari peluang karier baru di Malaysia yang menjanjikan kehidupan yang
   lebih baik. Kembangkan karakter dan latar belakang protagonis, serta ceritakan perjalanan
   dan pertimbangannya dalam mengambil keputusan penting tersebut.
2 Answer: Di sebuah bandar kecil di Indonesia, hiduplah seorang wanita bernama Maya. Maya
   adalah seorang pekerja keras yang telah bekerja di sebuah syarikat besar selama lima
   tahun. Dia mempunyai pekerjaan yang stabil dan gaji yang mencukupi. Namun, dalam hatinya,
   Maya merasakan bahawa pekerjaannya tidak memberikan kepuasan dan cabaran yang mencukupi
   untuknya. Maya bercita-cita untuk meningkatkan kerjaya dan mencapai kejayaan yang lebih
   besar. Dia sering bermimpi untuk hidup yang lebih baik, dengan pendapatan yang lebih
   tinggi dan peluang untuk mengembara ke seluruh dunia. Pada suatu hari, Maya mendengar
   berita bahawa di Malaysia terdapat peluang kerjaya yang menjanjikan. Maya menjadi
   tertarik dengan peluang ini. Dia mula mencipta tentang industri dan peluang pekerjaan di
   Malaysia. Maya tahu bahawa kerjaya di sana boleh memberikan pendapatan yang jauh lebih
   tinggi daripada yang dia terima sekarang. Selain itu, dia juga tahu bahawa di Malaysia
   terdapat persekitaran kerja yang lebih dinamik dan berbeza-beza, yang akan memberikan
   cabaran baru untuknya. Namun, Maya juga menyedari bahawa untuk mencari peluang kerjaya di
   Malaysia, dia harus meninggalkan semua yang dia telah bangukan di Indonesia, termasuk
   keluarga dan rakan-rakannya. Maya berasa sedih dan terfikir, kerana dia sangat menyayangi
   keluarga dan tidak tahan berpisah dengan mereka. Maya menghabiskan masa berhari-hari
   memikirkan pilihan kerjaya ini. Dia menulis segala kelebihan dan kekurangan kedua-dua
   pilihan yang dia hadapi. Dia bercakap dengan keluarga dan rakan-rakannya, dan mereka
   menyokong keputusannya, walaupun mereka sedih dengan pemergian Maya. Akhirnya, selepas
   pertimbangan yang teliti, Maya memutuskan untuk mengambil risiko dan mencari peluang
   kerjaya baru di Malaysia. Dia merasakan bahawa ini adalah peluang yang jarang berlaku,
   dan dia harus mengambil langkah di luar zon selesa untuk mencapai impian dan
   cita-citanya. Maya menyedari bahawa hidup adalah tentang mengambil risiko dan menghadapi
   cabaran baru. Perjalanan Maya ke Malaysia bukanlah mudah. Dia harus melalui beberapa
   halangan dan penyesuaian budaya baru. Namun, dengan ketabahan dan semangat yang kuat,
   Maya berjaya mendapatkan pekerjaan di sebuah syarikat besar di Kuala Lumpur. Dia berasa
   gembira dan bangga dengan keputusannya. Selama bertahun-tahun di Malaysia, Maya mencapai
   kejayaan yang luar biasa. Dia menikmati kerjaya yang dia lalui dan mempunyai kehidupan
   yang lebih baik dan lebih selesa daripada yang pernah dia bayangkan di Indonesia. Maya
   juga terus menjalin hubungan dengan keluarga dan rakan-rakannya di negara ini, walaupun
   jarak memisahkan mereka. Dalam cerita ini, Maya adalah contoh yang memberi inspirasi
   kepada seseorang yang berani mengambil risiko dengan meninggalkan pekerjaan yang stabil
   dan mencari peluang kerjaya baru. Keputusannya membawa perubahan besar dalam hidupnya,
```

dan ia menunjukkan kepada kita bahawa kadang-kadang kita harus berani keluar dari zon selesa untuk mencapai impian dan cita-cita kita.

All synthetic dataset and implementation published at [mesolitica/chatgpt-malay-instructions](https://mesolitica.com/chatgpt-malay-instructions).

4.1.3 Malay UltraChat

We extended our efforts to generate a multiturn dataset. Acknowledging the efficacy of multiturn conversations in training language models to understand context and coherence, we sought inspiration from Ultrachat [5]. Ultrachat served as a valuable reference, guiding us in crafting a dataset that incorporates multifaceted interactions and conversations.

```
1  [{ 'role': 'context',
2    'content': 'LIGA\xa0Malaysia (Liga-M) terus menerima tamparan hebat apabila UiTM FC sah
    menarik diri daripada saingan Liga Super musim ini bagi menyusuli jejak Petaling Jaya
    City.\n\n\n\nPerkara itu disahkan sendiri oleh Presiden The Lion Troops, Profesor Dr.
    Sazili Shahibi, selepas dua musim beraksi di pentas Liga Super.\n\n\n\nSelain masalah
    kewangan, Dr Sazili menjelaskan ada punca lain mengapa pihak pengurusan mengibarkan
    bendera putih.\n\n\n\n\n\n\n\nAhli Lembaga Pengarah UiTM telah bermesyuarat dan (kami
    mengambil) keputusan untuk menarik diri daripada liga pada musim
    depan.\n\n\n\n\n\n\n\n"Kewangan adalah salah faktornya, tapi kami rasa daripada segi kemampuan,
    ia tidak sesuai buat pemain-pemain kami," katanya ketika dihubungi \nAstro
    Arena\n\xa0pada Rabu.\n\n\n\n\n\n\n\nSazili juga berhasrat memaklumkan UiTM bakal beralih ke
    saingan Liga M3 musim ini.\n\n\n\n\n\n\n\nUntuk rekod, UiTM muncul pasukan berstatus universiti
    pertama yang layak beraksi dalam Liga Super.\n\n\n\n\n\n\n\nDalam pada itu, Liga Bolasepak
    Malaysia (MFLhanya akan menjelaskan situasi berkenaan menerusi kenyataan rasmi bakal
    dalam masa terdekat selepas hanya 14 pasukan yang akan bertanding di pentas Liga Super
    musim ini. -\nStadium Astro',
3    'content_ms': None,
4    'indon': False},
5  { 'role': 'user',
6    'content': 'What are the reasons given by UiTM FC for withdrawing from Liga Malaysia?',
7    'content_ms': 'Apakah alasan yang diberikan oleh UiTM FC kerana menarik diri daripada Liga
    Malaysia?',
8    'indon': False},
9  { 'role': 'assistant',
10   'content': 'Menurut Presiden UiTM FC, Profesor Dr. Sazili Shahibi, mereka mengambil keputusan
    untuk menarik diri daripada Liga Malaysia musim ini atas beberapa sebab. Salah satu sebab
    utamanya adalah masalah kewangan yang dihadapi oleh pasukan. Selain itu, beliau juga
    menyebut bahawa kemampuan pasukan UiTM tidak sesuai untuk bersaing dalam Liga Super.
    Selain masalah ini, tidak ada sebab lain yang diberikan dalam laporan tersebut.',
11   'content_ms': 'Menurut Presiden UiTM FC, Profesor Dr. Sazili Shahibi, mereka membuat
    keputusan untuk menarik diri daripada Liga Malaysia musim ini atas beberapa sebab. Salah
    satu sebab utama ialah masalah kewangan yang dihadapi oleh pasukan. Selain itu, beliau
    juga menyatakan bahawa keupayaan pasukan UiTM tidak sesuai untuk bersaing dalam Liga
    Super. Selain masalah ini, tiada sebab lain diberikan dalam laporan tersebut.',
12   'indon': False},
13 { 'role': 'user',
14   'content': 'Alamak, memang sedihlah tengok UiTM FC tarik diri dari Liga Malaysia. Macam mana
    pula dengan penonton setia yang selalu sokong pasukan tu? Harap-harap ada jalan
    penyelesaian untuk masalah kewangan mereka.',
15   'content_ms': 'Alamak, memang sedih tengok UiTM FC menarik diri dari Liga Malaysia. Bagaimana
    pula dengan penonton setia yang selalu menyokong pasukan itu? Harap-harap ada jalan
    penyelesaian untuk masalah kewangan mereka.',
16   'indon': False},
17 { 'role': 'assistant',
18   'content': 'Ya, memang sedih melihat UiTM FC menarik diri dari Liga Malaysia. Bagi penonton
    setia yang sentiasa menyokong pasukan tersebut, tentu mereka juga berasa kecewa. Kami
```


juga berharap ada jalan penyelesaian untuk masalah kewangan yang dihadapi oleh UiTM FC. Mungkin pihak pengurusan boleh mencari sponsor atau menyusun semula strategi kewangan untuk membolehkan pasukan tersebut terus beraksi dalam liga pada masa hadapan. Semoga mereka dapat menemui penyelesaian yang baik dan kembali bersaing di pentas bola sepak Malaysia.',

19 'content_ms': 'Ya, memang sedih melihat UiTM FC menarik diri dari Liga Malaysia. Bagi penonton setia yang sentiasa menyokong pasukan tersebut, tentu mereka juga berasa kecewa. Kami juga berharap ada jalan penyelesaian untuk masalah kewangan yang dihadapi oleh UiTM FC. Mungkin pihak pengurusan boleh mencari penaja atau menyusun semula strategi kewangan untuk membolehkan pasukan tersebut terus beraksi dalam liga pada masa hadapan. Semoga mereka dapat menemui penyelesaian yang baik dan kembali bersaing di pentas bola sepak Malaysia.',

20 'indon': False},

21 {'role': 'user',

22 'content': 'Ya betul tu, banyak peminat setia UiTM FC yang akan kecewa dengan keputusan mereka tarik diri dari Liga Malaysia. Harapnya pihak pengurusan dapat cari jalan penyelesaian secepat mungkin supaya pasukan tu masih dapat beraksi dalam liga. Saya pun tak sabar nak tengok prestasi UiTM FC lagi musim hadapan. Semoga ada berita baik dalam masa terdekat.',

23 'content_ms': 'Ya, betul tu. Banyak peminat setia UiTM FC yang akan kecewa dengan keputusan mereka menarik diri dari Liga Malaysia. Harapnya pihak pengurusan dapat mencari jalan penyelesaian secepat mungkin supaya pasukan itu masih dapat beraksi dalam liga. Saya pun tak sabar nak tengok prestasi UiTM FC lagi musim depan. Semoga ada berita baik dalam masa terdekat.',

24 'indon': False},

25 {'role': 'assistant',

26 'content': 'Saya juga setuju dengan anda. Semoga pihak pengurusan UiTM FC dapat mencari jalan penyelesaian secepat mungkin agar pasukan tersebut masih dapat berlaga dalam Liga Malaysia. Benar, banyak peminat yang ingin melihat prestasi UiTM FC musim hadapan. Mari kita nantikan berita baik daripada mereka dalam masa terdekat. Harapnya mereka dapat mengatasi masalah kewangan dan kembali membanggakan para penyokong. Kita terus sokong UiTM FC!',

27 'content_ms': 'Saya juga bersetuju dengan anda. Semoga pihak pengurusan UiTM FC dapat mencari jalan penyelesaian secepat mungkin supaya pasukan tersebut masih dapat bersaing dalam Liga Malaysia. Benar, ramai peminat yang ingin melihat prestasi UiTM FC musim depan. Mari kita nantikan berita baik daripada mereka dalam masa terdekat. Harapnya mereka dapat mengatasi masalah kewangan dan kembali membanggakan para penyokong. Kita terus sokong UiTM FC!',

28 'indon': False},

29 {'role': 'user',

30 'content': 'Ya lah, harapnya UiTM FC dapat dapat mencari penyelesaian yang cepat. Penyokong setia UiTM FC mesti excited nak tengok pasukan beraksi lagi musim hadapan. Semoga ada berita gembira dalam masa terdekat!',

31 'content_ms': 'Ya lah, harapnya UiTM FC dapat mencari penyelesaian yang cepat. Penyokong setia UiTM FC mesti teruja untuk melihat pasukan beraksi lagi musim depan. Semoga ada berita gembira dalam masa terdekat!',

32 'indon': False},

33 {'role': 'assistant',

34 'content': 'Betul tu, penyokong setia UiTM FC pasti sangat teruja melihat pasukan beraksi di musim hadapan. Kita harap pihak pengurusan dapat mencari penyelesaian dengan pantas supaya pasukan dapat kembali ke gelanggang. Kita tunggu berita gembira yang akan datang dan terus menyokong UiTM FC dengan penuh semangat. Biar bola sepak Malaysia sentiasa berwarna dengan kehadiran pasukan hebat seperti UiTM FC!',

35 'content_ms': 'Betul tu, penyokong setia UiTM FC pasti sangat teruja melihat pasukan beraksi pada musim depan. Kita harap pihak pengurusan dapat mencari penyelesaian dengan pantas supaya pasukan dapat kembali ke gelanggang. Kita tunggu berita gembira yang akan datang dan terus menyokong UiTM FC dengan penuh semangat. Biar bola sepak Malaysia sentiasa

```

berwarna dengan kehadiran pasukan hebat seperti UiTM FC!',
36 'indon': False},
37 {'role': 'user',
38 'content': 'Ya betul tu, memang berharap pasukan UiTM FC dapat segera mencari jalan
penyelesaian. Setiap pasukan memang perlu sokongan dan kami sebagai penyokong setia akan
terus memberikan sokongan kepada pasukan kesayangan kami. Semoga ada berita gembira yang
akan datang!'},
39 'content_ms': 'Ya, betul tu. Memang berharap pasukan UiTM FC dapat segera mencari jalan
penyelesaian. Setiap pasukan memang perlukan sokongan dan kami sebagai penyokong setia
akan terus memberikan sokongan kepada pasukan kesayangan kami. Semoga ada berita gembira
yang akan datang!'},
40 'indon': False},
41 {'role': 'assistant',
42 'content': 'Setiap pasukan memang memerlukan sokongan daripada penyokong setia seperti anda.
Anda adalah aset berharga bagi pasukan UiTM FC. Teruskan menyokong dengan penuh semangat
dan harapan. Kita berdoa agar pihak pengurusan dapat mencari jalan penyelesaian dengan
segera. Bersama-sama, mari kita nantikan berita gembira yang akan datang dan terus
memberikan sokongan kepada pasukan kesayangan kita. UiTM FC pasti akan bangkit dan
menyampaikan prestasi yang cemerlang musim hadapan!'},
43 'content_ms': 'Setiap pasukan memang memerlukan sokongan daripada penyokong setia seperti
anda. Anda adalah aset berharga bagi pasukan UiTM FC. Teruskan menyokong dengan penuh
semangat dan harapan. Kita berdoa agar pihak pengurusan dapat mencari jalan penyelesaian
dengan segera. Bersama-sama, mari kita nantikan berita gembira yang akan datang dan terus
memberikan sokongan kepada pasukan kesayangan kita. UiTM FC pasti akan bangkit dan
menyampaikan prestasi yang cemerlang musim depan!'},
44 'indon': False}]

```

All synthetic dataset and implementation published at [mesolitica/malaysian-ultrachat](https://mesolitica.com/malaysian-ultrachat).

4.1.4 Synthetic Malay CommonSense

include some data.

Complete generating synthetic dataset implementation at [here](#).

4.1.5 Coding Dataset

include some data.

Complete generating synthetic dataset implementation at [here](#).

4.2 Malay Instruction Dataset

4.2.1 Synthetic Kesalahan Tatabahasa Dataset

Complete generating synthetic dataset implementation at [here](#).

4.2.2 Synthetic Noisy Stemming Dataset

Complete generating synthetic dataset implementation at [here](#).

4.2.3 Dewan Bahasa Pustaka Definition Dataset

Complete generating synthetic dataset implementation at [here](#).

4.3 Finetuning Phase

Use Mistral chat template, mention the chat template below,

DeepSpeed Zero3, batch size 6, packing, 16384 context length, constant learning rate 2e-5.

4.4 Result Finetuning

4.4.1 Multiturn Malaysian context QA

4.4.2 Multiturn Coding QA

4.4.3 Translation low language

4.4.4 Malay instruction

5 Evaluation

We use Tatabahasa dataset, gathered from <https://tatabahasabm.tripod.com/latih/latih.htm>, contain 349 questions.

We published at <https://huggingface.co/spaces/mesolitica/malay-llm-leaderboard>

6 Acknowledgement

Special thanks to Malaysia-AI volunteers.

7 Conclusion

able to reduce research gap.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
- [2] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023.
- [3] Zolkepli Husein. Malaya, natural-language-toolkit library for bahasa malaysia, powered by pytorch. <https://github.com/huseinzol05/malaya>, 2018.
- [4] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023.
- [5] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.