



Kertas Kerja

LLM – Bahasa Melayu Rasmi

2024-11-01

DISEDIKAN OLEH:

Vince Ng, Balraj Pannu and Wan M Hasni Wan Sulaiman

DWH Mesra AI | DWH Technology

DWH Mesra AI Sdn. Bhd., DWH Technology, Suite 19-12 G Tower Kuala Lumpur, 50450 Kuala Lumpur, Wilayah



Persekutuan Kuala Lumpur.

- www.dwh-international.com

Isi Kandungan

1. Ringkasan Eksekutif	2
2. Objektif dan Matlamat LLM	2
2.1 Objektif	2
2.2 Matlamat.....	2
3. Pengenalan	3
4. Definisi “Large Language Models”	3
5. Perkembangan LLM	4
6. Pembelajaran LLM Memerlukan Jumlah Data Yang Besar	6
7. Pembelajaran LLM Memerlukan Sumber Perkomputeran Yang Tinggi.....	7
8. Keterbatasan Pada LLM Yang Sedia Ada	8
9. Permasalahan Bahasa Melayu Dalam LLM	9
9.1 Status Bahasa Melayu Di Dalam LLM Sedia Ada	9
9.2 Ketersediaan Korpus Bahasa Melayu Sebagai Sumber Data Pembelajaran LLM	11
10. Ujian Terhadap LLM Sedia Ada Mengenai Bahasa Melayu	13
10.1 Ujian Maksud Perkataan.....	13
10.2 Terjemahan	15
10.3 Ringkasan Atau Kesimpulan	18
11. Usaha-Usaha Yang Sedia Ada Untuk Membina LLM Bahasa Melayu.....	19
11.1 Projek Malaya oleh Mesolitica.....	19
11.2 Projek MelayuBERT	20
12. Kenapa Perlu Membina LLM Bahasa Melayu Rasmi?	21

Maklumat yang diberikan dalam cadangan ini adalah peribadi dan sulit; oleh itu, sesiapa yang membaca cadangan ini bersetuju untuk tidak mendedahkan sebarang maklumat di dalamnya tanpa kebenaran bertulis daripada Syarikat terlebih dahulu.

1. Ringkasan Eksekutif

Projek ini bertujuan untuk membangunkan “LLM” Bahasa Melayu Rasmi (BMR) menggunakan teknologi “Large Language Model” (LLM) yang memenuhi piawaian bahasa Melayu formal di Malaysia. Dengan perkembangan pesat penggunaan AI dalam kehidupan harian rakyat Malaysia, terdapat keperluan kritikal untuk membangunkan model bahasa yang mampu memahami dan menghasilkan Bahasa Melayu Rasmi dengan ketepatan dan ketulenan budaya. Model bahasa sedia ada sering tidak memenuhi piawaian bahasa rasmi dan mengandungi bias terhadap Bahasa Indonesia. Hal ini menjadikan penghasilan BMR satu keutamaan nasional.

Projek ini akan dilaksanakan melalui kerjasama erat antara Dewan Bahasa dan Pustaka (DBP) dan pakar AI dari DWH. Sumber data rasmi seperti dokumen DBP, penerbitan akademik, dan korpus bahasa akan digunakan untuk memastikan ketepatan dan ketulenan model. Selain itu, infrastruktur pengkomputeran berprestasi tinggi, termasuk kluster GPU dan perkhidmatan ‘awan’ akan dilaburkan bagi menyokong latihan model berskala besar.

Model Bahasa Melayu Rasmi ini bukan sahaja akan memelihara dan mempromosikan penggunaan Bahasa Melayu dalam teknologi terkini tetapi juga akan memberi impak besar kepada sektor kerajaan, pendidikan, dan perkhidmatan awam. Dengan pelaksanaan dan pengkomersialan model ini, DBP dapat mengambil peranan utama dalam transformasi digital negara, menjana pendapatan melalui produk berasaskan AI, dan menyumbang kepada pembangunan ekonomi digital Malaysia.

2. Objektif dan Matlamat LLM

2.1 Objektif

1. **Membangunkan LLM-BMR:** Membina model bahasa AI yang dilatih menggunakan data rasmi untuk memenuhi piawaian bahasa formal Malaysia.
2. **Mengatasi Bias Bahasa:** Mengurangkan bias terhadap Bahasa Indonesia dalam model sedia ada dengan menggunakan sumber data yang berkualiti tinggi dan khusus untuk Bahasa Melayu.
3. **Menyokong Transformasi Digital:** Menyediakan alat AI yang boleh digunakan dalam sektor kerajaan, pendidikan, dan perkhidmatan awam bagi mempertingkatkan kecekapan dan ketepatan komunikasi.

2.2 Matlamat

1. **Ketepatan dan Ketulenan Bahasa:** Memastikan model AI dapat memahami dan menghasilkan Bahasa Melayu Rasmi dengan ketepatan tinggi.
2. **Penggunaan Meluas dalam Pelbagai Sektor:** Meningkatkan penggunaan BMR dalam sektor kerajaan, pendidikan, dan komersial.
3. **Kepimpinan dalam Teknologi AI Tempatan:** Menjadikan Malaysia peneraju dalam teknologi AI berasaskan Bahasa Melayu, menyumbang kepada ekonomi digital negara.

3. Pengenalan

Kertas kerja ini bertujuan untuk menyatakan cadangan kerja untuk membina suatu model yang dinamakan sebagai **LLM-BMR** ia akan diusahakan sebagai satu kerjasama di antara Dewan Bahasa dan Pustaka (DBP) dan DWH Mesra AI Sdn. Bhd. Kaedah dan bentuk kerjasama akan diperjelaskan di dalam kertas kerja ini, dan ia akan tertakluk kepada perjanjian yang bakal dibuat di antara kedua pihak, setelah mendapat persetujuan dasar dari pihak masing-masing.

Di dalam kertas kerja ini, kami akan merangka apa yang dimaksudkan dengan *Large Language Models (LLM)*, dan bagaimanakah ia diaplikasikan dalam dunia digital pada hari ini. Seterusnya kami juga akan memperjelaskan kenapa **LLM-BMR** adalah satu perkara yang penting untuk dibina. Kami juga memperincikan kenapa pada masa kini, LLM yang ada mempunyai permasalahan yang agak ketara di mana Bahasa Melayu yang terbentuk di dalamnya terpesong daripada bahasa yang sebenar, terutamanya disebabkan kemasukan atau percampuran di antara Bahasa Melayu dengan Bahasa Indonesia.

Seterusnya kami akan mencadangkan langkah-langkah serta kaedah-kaedah bagaimana **LLM-BMR** ini boleh dibina. Ini termasuklah strategi pengumpulan data (dalam bentuk korpus serta repositori), pembuatan model LLM, kaedah latihan model, serta kaedah penilaian penanda aras (*benchmarking*) yang perlu dilaksanakan.

4. Definisi “Large Language Models”

Mengikut kamus Cambridge, “*Large Language Models*” (LLM) adalah “*a complex mathematical representation of language that is based on very large amounts of data and allows computers to produce language that seems similar to what a human might say*”. Ia diterjemahkan melalui ChatGPT sebagai: “Suatu representasi matematik yang kompleks tentang bahasa berdasarkan sejumlah besar data dan membolehkan komputer menghasilkan bahasa yang kelihatan serupa dengan apa yang mungkin dikatakan oleh manusia.”

Dalam Bahasa Melayu pula, carian yang kami buat melalui Kamus Dalam Talian DBP menunjukkan tiada maklumat yang ada. Ia diterjemahkan secara langsung sebagai “Model Bahasa Besar” atau “Model Bahasa Skala Besar”. Kami lebih bersetuju kepada “Model Bahasa Skala Besar” memandangkan “*Large Scale*”, yang membawa makna “Skala Besar”.

Penjelasan yang paling tepat dibuat oleh Stephen Wolfram di dalam buku nya “*What is ChatGPT Doing... and Why Does It work?*”. Beliau menjelaskan bahawa ChatGPT (dan juga mana-mana model GPT) adalah merupakan satu program komputer yang memberikan “*reasonable continuation*” (kesinambungan yang munasabah) kepada apa-apa teks yang dimasukkan ke dalam program tersebut. Bagaimana ia dapat menentukan kesinambungan yang munasabah, adalah disebabkan “ia telah melihat hal yang sama pada berbilion halaman web dan lain-lain teks”.

Secara teknikal pula, LLM boleh dikatakan sebagai perantara di antara manusia dengan komputer. Manusia memasukkan soalan atau memberi tugas kepada komputer dengan menggunakan bahasa biasa (*natural language*). Komputer memahami maksud soalan atau tugas tersebut, dan terus melaksanakan pelbagai kod komputer untuk menjawab atau memenuhi tugas tersebut. Hasil daripada soalan tersebut, atau tugas tersebut akan difahami oleh penyوال atau pemberi tugas dengan mudah dalam bahasa biasa. Ini menjadikan LLM sebagai suatu yang dinamakan kecerdasan buatan atau *Artificial Intelligence*.

Dengan adanya LLM, seseorang tidak lagi perlu kepada pengetahuan untuk menulis kod komputer, atau mahir dalam bahasa peranti atau bahasa pengaturcaraan komputer. Sebaliknya, apa diperlukan ialah menggunakan bahasa biasa (*natural language*), dengan kaedah yang mudah tanpa perlu kepada ketepatan tatabahasa, bahkan tanpa perlu ketepatan kepada ejaan dan terma yang digunakan.

5. Perkembangan LLM

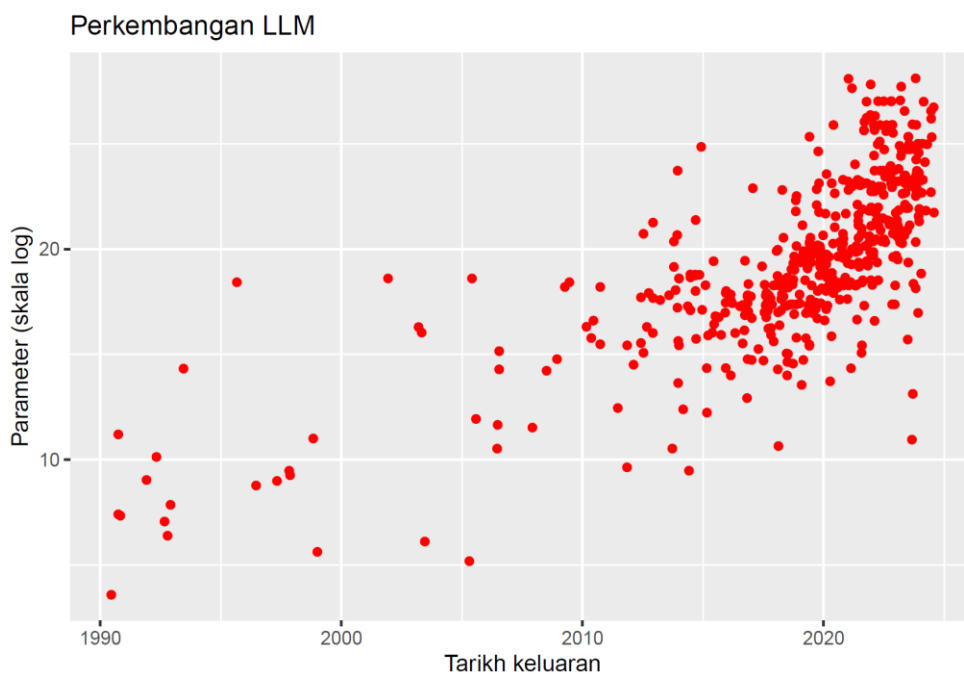
Permulaan awal perkembangan LLM adalah dalam bidang pengkhususan “*Natural Language Processing*” (NLP), iaitu “Pemprosesan Bahasa Semula Jadi” dan juga dalam aspek “*Computational Linguistics*” (CL), iaitu “Pengkomputeran Linguistik”. Namun perkembangan permodelan bahasa mengikut kaedah NLP dan CL agak perlahan disebabkan kerumitan menterjemahkan “*rules of human language*” kepada “*rules of computing*”. Terutamanya apabila tatabahasa serta lenggok bahasa seperti persoalan medan makna bahasa (*semantics*) serta konteks penggunaan terma-terma tertentu dalam sesuatu bahasa sukar diterjemahkan kepada “*rules based computations*” dalam program pengkomputeran.

Disebabkan kerumitan tersebut, pada peringkat awal pembangunan NLP dan CL adalah berfokuskan kepada kaedah pembelajaran terarah (*supervised-learning*), di mana ia memerlukan penglabelan terhadap data-data yang digunakan sebagai bahan latihan (*training data*). Ketidacukupan data untuk latihan menjadi suatu (*insufficiency of training and testing data*) menyebabkan pembinaan model yang berkesan tidak dapat dihasilkan dengan cara mudah.

Perkara ini berubah apabila “*machine deep learning*” (pembelajaran mesin-mendalam) diperkenalkan melalui teknologi rangkaian neural (*neural networks*) berserta dengan teknologi pemkomputeran selari (*parallel computing*) menggunakan *Graphics Processing Units* (GPU). Ia mula diperkenalkan oleh Google melalui pakej Tensorflow (tahun 2015), yang kemudiannya dikembangkan kepada pelbagai pakej pengkomputeran yang lain.

Seterusnya ia dikembangkan kepada teknologi baru yang dinamakan sebagai “*transformer models*” (oleh Google) pada tahun 2017. Model *transformers* seterusnya dikembangkan kepada model-model yang dinamakan sebagai “*Generative Pre-trained Transformers*” (GPT). Inilah merupakan model asas untuk ChatGPT yang diperkenalkan oleh OpenAI (GPT1.0 atau ChatGPT) pada November 2022. ChatGPT oleh OpenAI bukanlah satu-satunya model *transformers* yang diperkenalkan, bahkan kini sudah wujud hampir 800 buah model besar yang sudah diperkenalkan ke pasaran.

Gambar rajah/chart di bawah memberikan ringkasan perkembangan LLM dari tahun 1990 sehingga Julai 2024.



Sumber data: <https://epochai.org/data/notable-ai-models>

Model berskala besar adalah di atas tanda aras 15 (jumlah parameter di atas 100 juta). Perkembangan pesat terjadi setelah “*transformers*” diperkenalkan pada tahun 2017 di mana pelbagai jenis model telah ditampilkan ke dalam pasaran.

Jadual berikut menyenaraikan beberapa model besar sebagai rujukan.

Sistem	Organisasi	Tarikh	Parameter
Llama 3.1-405B	Meta AI	2024-07-23	4.0500e+11
Nemotron-4 340B	NVIDIA	2024-06-14	3.4000e+11
MegaScale (Production)	ByteDance, Peking University	2024-02-23	5.3000e+11
QMoE: compressed 1T model	Institute of Science and Technology Austria (ISTA), Neural Magic	2023-10-25	1.6000e+12
PaLM 2	Google	2023-05-10	3.4000e+11
PanGu- Σ	Huawei Noah's Ark Lab	2023-03-20	1.0850e+12
PaLM-E	Google, TU Berlin	2023-03-06	5.6200e+11
U-PaLM (540B)	Google	2022-10-20	5.4000e+11
LMSI-Palm	Google, University of Illinois Urbana-Champaign (UIUC)	2022-10-20	5.4000e+11
Flan-PaLM 540B	Google	2022-10-20	5.4000e+11
Minerva (540B)	Google	2022-06-29	5.4035e+11
PaLM (540B)	Google Research	2022-04-04	5.4035e+11
GLaM	Google	2021-12-13	1.2000e+12
Megatron-Turing NLG 530B	Microsoft, NVIDIA	2021-10-11	5.3000e+11
M6-T	Alibaba	2021-03-05	1.0027e+12
Switch	Google	2021-01-11	1.5710e+12

Sumber data: <https://epochai.org/data/notable-ai-models>

Secara umumnya, trend menunjukkan model-model LLM mutakhir menggunakan jumlah parameter yang cukup besar. Sebagai contoh, Llama3.1-405B, menggunakan 405 billion parameter. QMoE pula menggunakan 1.6 trilion parameter. Hampir kesemua model-model LLM terkini sudah berada pada tahap 300 bilion parameter ke atas. Ini adalah jauh lebih besar daripada model-model asas, seperti GPT2, BERT, Llama 2, dan sebagainya, di mana rata-rata jumlah parameter hanyalah bersekitaran 50 bilion.

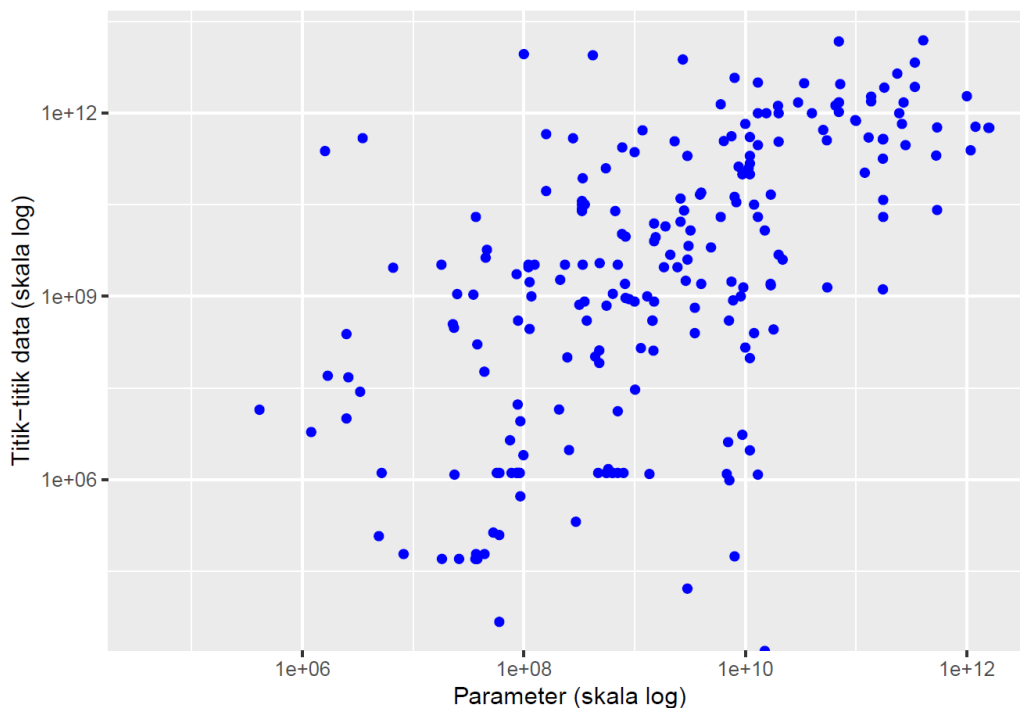
6. Pembelajaran LLM Memerlukan Jumlah Data Yang Besar

Sebagaimana yang dijelaskan oleh Stephen Wolfram, LLM merupakan satu “*imprint*” kepada semua maklumat yang ada di dalam keseluruhan dunia maya (Internet). Cetakan tersebut perlu disimpan di dalam parameter model, sebagai pemberat model dalam angka-angka kebarangkalian. Untuk memastikan data-data tersebut diserapkan ke dalam model, ia perlu di“lihat” berulang-ulang kali (berbillion kali), supaya ia menjadi “*hard wired*” kepada model tersebut. Ini membawa kepada keperluan kerencaman dalam data, kepelbagaian kepada data, dan sebagainya supaya kesamaan dalam data diminimakan dengan sebaik mungkin.

Sebagai maklumat umum juga, sekiranya model mempunyai parameter yang banyak, model tersebut boleh menjalankan tugas yang lebih besar. Sebagai contoh, model berparameter 2 billion, mungkin boleh mengambil input yang panjang (sebagai contoh 2000 token) dan mengeluarkan output yang panjang (sebagai contoh 6000 token). Untuk mutakhir ini, model-model berparameter besar sudah boleh mengambil input 10,000 token, dan menghasilkan output sebanyak 128,000 token. Model berparameter besar juga boleh melaksanakan tugas yang lebih rumit serta kompleks.

Untuk membina model berparameter besar, ia memerlukan jumlah data yang besar. Justeru, mana-mana LLM terkini kesemuanya memerlukan jumlah data yang besar untuk model dilatih dengan sempurna. Jumlah data dalam pembelajaran LLM diukur melalui jumlah token yang diguna pakai dan jumlah “parameter peringkat atas” (*hyperparameters*) ataupun “parameter” kepada model tersebut. Kadar yang boleh diambil sebagai ukuran ialah titik-titik data (*data points*) daripada jumlah parameter. Pada masa kini, kadarnya ialah satu titik data untuk satu parameter.

Carta di bawah menyenaraikan model, jumlah titik-titik data yang digunakan sebagai data pemula dan jumlah parameter.



Sumber data: <https://epochai.org/data/notable-ai-models>

Sebagaimana tertera dalam graf diatas, model yang mempunyai 1 trilion parameter memerlukan 1 trilion titik-titik data. Oleh yang demikian, untuk sesiapa yang mahu membuat model LLM yang baru daripada awal, perlu mengambil kira jumlah data yang ada dan juga saiz model (parameter) yang perlu untuk memastikan model tersebut boleh membuat tugas yang dikehendaki.

7. Pembelajaran LLM Memerlukan Sumber Perkomputeran Yang Tinggi

Tiga faktor utama kenapa LLM memerlukan sumber perkomputeran yang tinggi (*high performance computing*):

1. Saiz model LLM adalah besar (mengikut jumlah parameter).
2. Saiz data untuk latihan pembelajaran juga besar (mengikut jumlah token atau titik-titik data).
3. Kaedah pengkomputeran yang memerlukan banyak kiraan matematik (*massively computative*).

Model *transformers* dan juga mana-mana “*generative models*”, adalah merupakan satu mesin kiraan yang besar. Ia melibatkan kiraan algebra yang banyak, terutamanya dalam bentuk vektor dan juga pendaraban matrik, serta teknik pengoptimuman yang kompleks.

Sistem	Perkakasan	Jumlah	FLOP	Jam	Parameter
Llama 3.1-405B	NVIDIA H100 SXM5 80GB	16000	3.8000e+25	2142	4.0500e+11
Nemotron-4 340B	NVIDIA H100 SXM5 80GB	NA	1.8000e+25	2200	3.4000e+11
MegaScale (Production)	NVIDIA A100	12288	1.2000e+25	504	5.3000e+11
QMoE: compressed 1T model	NVIDIA RTX A6000	NA	NA	NA	1.6000e+12
PaLM 2	Google TPU v4	NA	7.3400e+24	NA	3.4000e+11
PanGu-Σ	Huawei Ascend 910	512	4.6700e+23	2400	1.0850e+12
PaLM-E	NA	NA	NA	NA	5.6200e+11
U-PaLM (540B)	Google TPU v4	512	2.5300e+24	120	5.4000e+11
LMSI-Palm	NA	NA	NA	NA	5.4000e+11
Flan-PaLM 540B	Google TPU v4	512	2.5000e+24	37	5.4000e+11
Minerva (540B)	Google TPU v4	1024	2.7415e+24	696	5.4035e+11
PaLM (540B)	Google TPU v4	6144	2.5272e+24	1536	5.4035e+11
GLaM	Google TPU v4	1024	3.7400e+23	1366	1.2000e+12
Megatron-Turing NLG 530B	NVIDIA A100 SXM4 80 GB	4480	1.1700e+24	770	5.3000e+11
M6-T	NVIDIA Tesla V100 DGXS 32 GB	480	5.5000e+21	NA	1.0027e+12
Switch	Google TPU v3	1024	8.2200e+22	648	1.5710e+12

Kos untuk melatih model bergantung kepada perkakasan (*hardware*) yang diguna pakai, serta jumlah perkakasan, dan FLOP (*Floating Operating Points*) yang didarabkan dengan jumlah jam untuk latihan. Secara umumnya, sekiranya kita menggunakan “*cloud computing*” yang disewa, ia akan melibatkan jumlah belanja yang besar. Mengikut kiraan terkini, untuk melatih model seperti Flan-PaLM 540B (seperti dalam jadual di atas), ia akan menelan belanja tidak kurang daripada 50 juta Dolar Amerika (atau Ringgit Malaysia 250 juta). Kos pembuatan model yang berskala besar, menggunakan data yang besar, sudah tidak lagi mengambil kira ‘kos berpatutan’.

8. Keterbatasan Pada LLM Yang Sedia Ada

Beberapa keterbatasan atau had yang ada kepada LLM yang tersedia perlu diperjelaskan supaya tidak ada kesalahfahaman serta kekeliruan yang akan terjadi. Di sini kami perjelaskan beberapa keterbatasan yang diketahui umum untuk makluman.

1. LLM bukanlah satu “permodelan bahasa” (*language modeling*), iaitu ia bukan dibentuk sebagai membuat atau membina sesuatu model untuk sesuatu bahasa. Ia adalah satu permodelan membuat representasi dalam sesuatu bahasa manusia yang boleh difahami oleh komputer untuk komputer tersebut membuat sesuatu tugas, di mana hasil tugas tersebut boleh difahami oleh orang yang memberikan arahan/tugas tersebut.
2. Jenis-jenis tugas yang boleh dilaksanakan oleh LLM adalah terhad dan bergantung penuh kepada asas model tersebut. Ia juga terhad kepada apa yang diajarkan kepada model tersebut, terutamanya data-data pembelajaran (training data) yang digunakan sewaktu pembelajaran model (model training) dibuat.
3. Walaupun “model *fine-tuning*” (proses pembetulan model) dibuat menggunakan data-data yang tepat atau baharu, ianya tidak menjamin pembetulan tersebut dapat dibuat dengan sempurna. Bahkan, seringkali proses pembetulan “*fine-tuning*” menemui jalan buntu yang menyebabkan proses permodelan perlu dimulakan kembali dari awal (*from scratch*).
4. Masalah halusinasi (*hallucinations*) dan keterpesongan (*drifts*) didalam hampir kesemua LLM yang wujud pada hari ini terlalu besar dan membawa risiko tinggi kepada pengguna.

Sebagaimana yang dijelaskan sebelum ini, permasalahan LLM bukanlah semata-mata permasalahan “bahasa”, tetapi ia adalah merupakan “bahasa pengarah baru” (*new instructions language*) yang kini sudah mula digunakan secara meluas, terutamanya di dalam dunia digital hari ini. Kepenggunaan LLM sudahpun meluas dalam bentuk ChatGPT dan lain-lain sepertinya sebagai “pembantu” (*assistant*) dalam pelbagai tugas. Sebagai contoh, ia sudah mula digunakan secara meluas dalam sistem kesihatan, perundangan, pengurusan, pendidikan dan juga keselamatan negara (*national security*).

Satu lagi persoalan yang timbul ialah adakah mungkin kita membina model LLM pada skala kecil, seperti model yang mempunyai jumlah parameter yang tidak terlalu besar. Permasalahannya di sini ialah LLM tidak berfungsi dengan baik tanpa mempunyai jumlah parameter yang banyak. Sebagai contoh, untuk Bahasa Inggeris, LLM yang mempunyai keupayaan yang baik, seperti BERT-base (daripada Google), mempunyai 110 juta parameter, dan BERT-Large mempunyai 440 juta parameter. Jumlah sumber mula data untuk BERT-base adalah sebanyak 3.3 bilion token yang digunakan. Saiz data pemula dan jumlah parameter kepada model adalah suatu yang sejajar. Untuk sesuatu model menjadi sempurna, ia perlu dibina dengan jumlah parameter yang besar (di atas 100 juta), sementara untuk model boleh dilatih dengan sempurna, ia memerlukan jumlah data yang besar. *Ratio* parameter dengan data adalah di atas 3 bilion token/100 juta parameter. Model LLM mutakhir menggunakan *ratio* yang jauh lebih tinggi di mana *ratio* trilion token/billion parameter diguna pakai (seperti model Llama3.1:405B).

Keterbatasan pengkomputeran berprestasi tinggi (*high performance computing*) juga menjadi satu halangan besar, di mana untuk menjalankan model LLM ia memerlukan kepada komputer peribadi yang mempunyai kapasiti serta cip komputer yang canggih dan terkini. Komputer juga memerlukan Unit Pemprosesan Grafik (GPU) yang berkuasa tinggi, di mana kos nya agak tinggi dan tidak mudah mampu milik. Lantaran itu, kebanyakan aplikasi perlu dijalankan melalui API (*Application Programming Interface*) melalui pengkomputeran awan. Kos pengkomputeran awan juga adalah suatu yang tidak mudah mampu milik.

9. Permasalahan Bahasa Melayu Dalam LLM

9.1 Status Bahasa Melayu Di Dalam LLM Sedia Ada

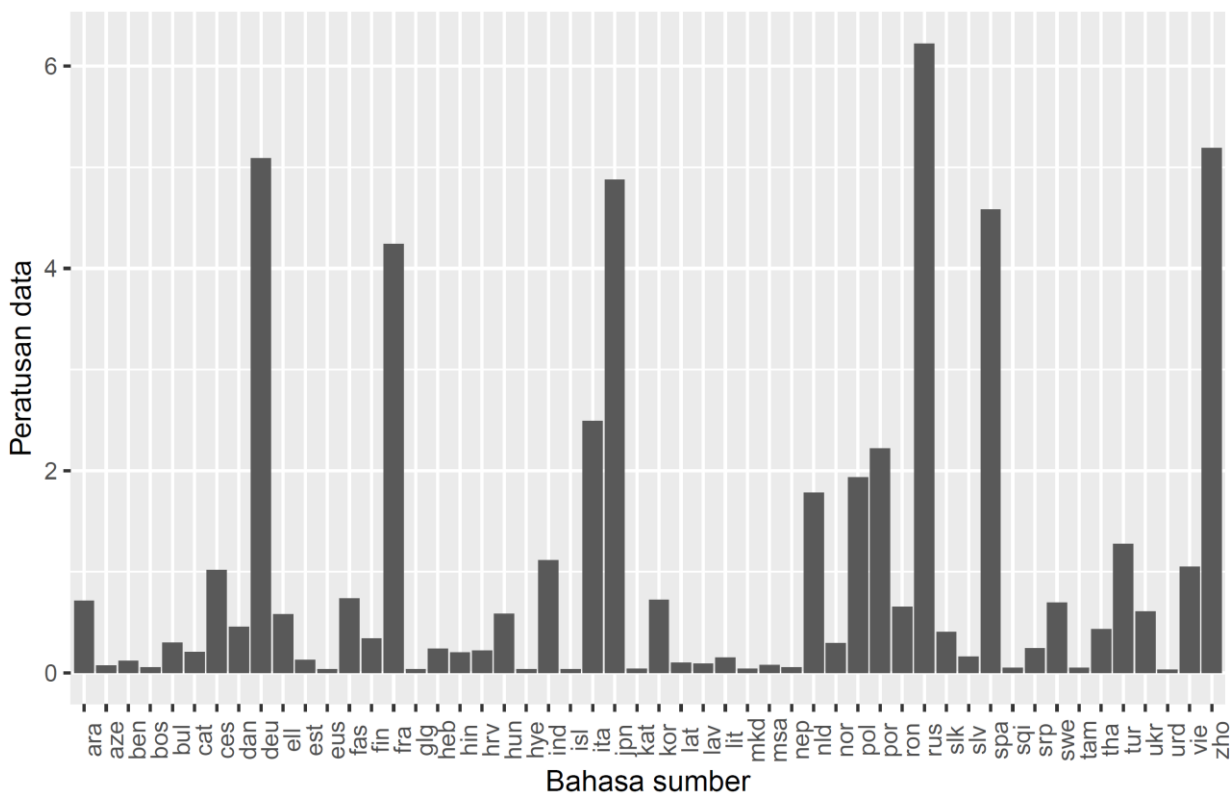
Data-data pembelajaran untuk Bahasa Melayu yang digunakan untuk melatih kebanyakan LLM adalah merupakan dari sumber data terbuka (*open source*). Terutamanya untuk data daripada Internet, ia diperolehi daripada Collosal Cleaned Crawled Corpus Organization atau ringkasannya C4. Di antara data yang banyak digunakan untuk bahasa-bahasa tertentu, diambil daripada Wikipedia Dump, atau Wikidump. Sebahagian organisasi menggunakan data terkumpul daripada platform-platform mereka seperti Facebook, Reddit, Github, dan sebagainya.

Sumber data C4

Data C4 adalah merupakan sedutan data daripada keseluruhan rangkaian web yang sedia ada. Proses pengklasifikasian sumber bahasa (*language source determination and classifications*) dibuat secara teratur, namun terdapat permasalahan besar di mana, ia gagal (pada kadar 40%) untuk membezakan sumber tersebut adalah Bahasa Melayu atau Bahasa Indonesia. Ia juga tidak membuat saringan terhadap penggunaan bahasa “pasar”, ataupun “loghat” dan sebagainya. Oleh sebab itu, kita tidak dapat membuat kepastian bahawa peratusan data-data tersebut memenuhi piawaian bahasa Melayu melebihi tahapan yang memadai.

Penjelasan daripada C4 untuk jumlah data-data daripada Bahasa Melayu yang digunakan sebagai data pembelajaran, daripada Bahasa Indonesia, adalah pada kadar 1:10. Ini memberikan impak besar kepada LLM yang dilatih, kerana dalam proses pengoptimuman model, ia akan lebih terikat kepada jumlah data yang lebih besar. Ini adalah berasaskan kepada pemodelan matematik kepada LLM yang dijelaskan sebelumnya.

Peratusan data C4 berdasarkan sumber bahasa (kecuali English)

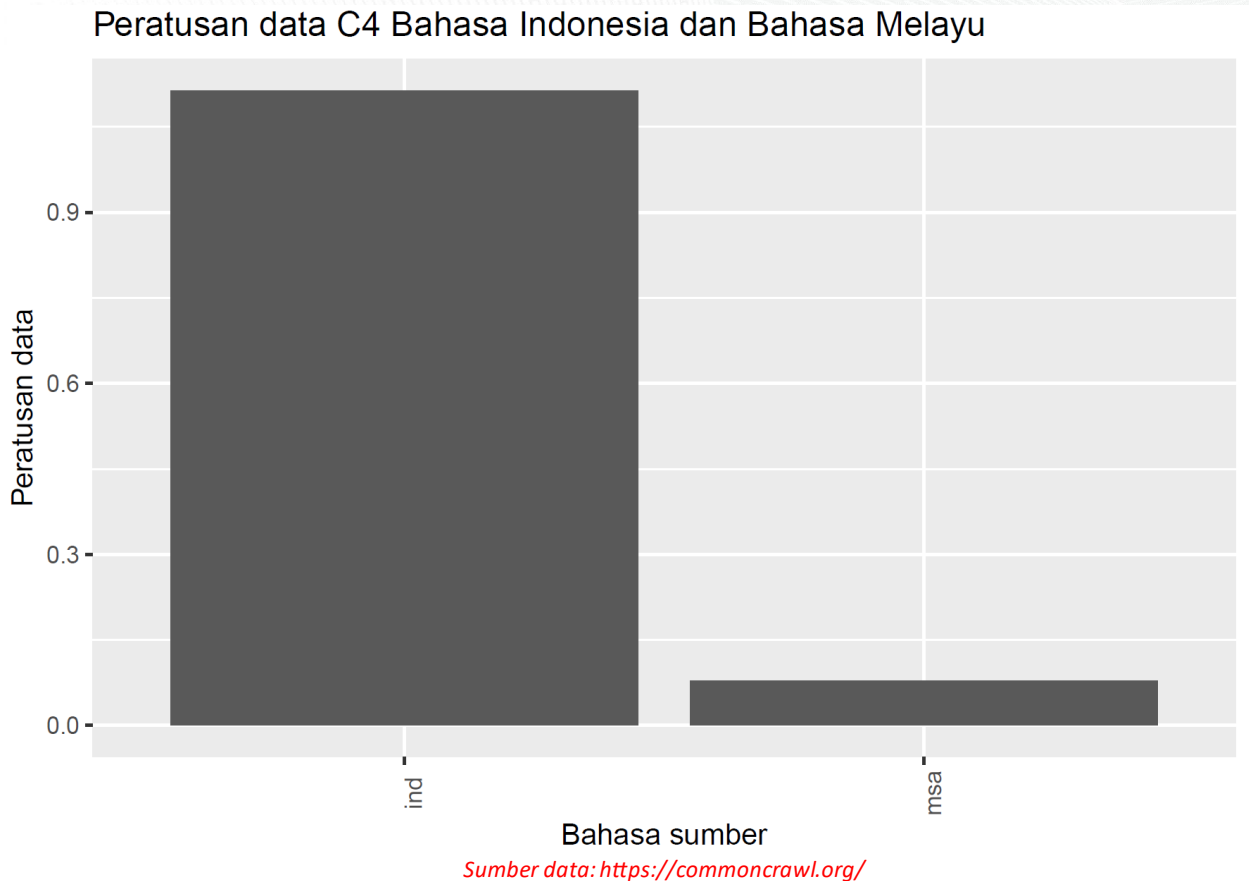


Sumber data: <https://commoncrawl.org/>

Dalam carta di atas, Bahasa Melayu memakai kod “msa”, sementara Bahasa Indonesia memakai kod “ind”.

Komposisi Bahasa Melayu di dalam data C4 hanyalah 0.0782 peratus, sementara Bahasa Indonesia pula 1.11 peratus, *bahasa Inggeris* sebanyak 43.11 peratus, bahasa Cina 5.19 peratus. Apabila digabungkan Bahasa Indonesia dan Bahasa

Melayu, jumlahnya adalah di atas 1 peratus. Mana-mana bahasa di mana jumlah data yang kurang daripada 1 peratus adalah dikategorikan sebagai “bahasa sumber rendah”. Justeru itu, untuk latihan pembelajaran LLM, Bahasa Melayu tergolong di dalam kategori “bahasa sumber rendah”, maka ia memerlukan penggabungan dengan Bahasa Indonesia (atau pun sebaliknya, Bahasa Melayu ditambahkan ke dalam Bahasa Indonesia).



Perincian dalam Carta di atas menunjukkan bahawa Bahasa Melayu di dalam data C4, hanyalah merupakan 7 peratus sahaja daripada gabungan data melalui augmentasi kedua-dua sumber data tersebut.

Dalam aspek jumlah token pula, Bahasa Melayu daripada C4 mempunyai hampir 312.4 juta token dan Bahasa Indonesia pula mempunyai hampir 4.44 bilion token, manakala 174 bilion token dalam Bahasa Inggeris. Lantaran itu, untuk membina LLM berparameter besar berasaskan data daripada Bahasa Melayu Sahaja adalah amat sukar kerana kekurangan data. Walaupun digabungkan dengan Bahasa Indonesia sekalipun, ia masih mempunyai kekurangan yang agak ketara.

Sumber data Wikipedia Bahasa Melayu

Sumber data daripada Wikipedia Bahasa Melayu boleh diperolehi daripada wikidump untuk Bahasa Melayu seperti berikut (untuk “dump” yang dibuat pada 24 Ogos 2024): <https://dumps.wikimedia.org/mswiki/20240820/>. Sebagai contoh, saiz data yang belum diproses adalah 352.2 MB. Wikidump akan melaksanakan proses “dumping” sebanyak dua kali sebulan.

Sebagai perbandingan, wikidump untuk Wikipedia Bahasa Indonesia pula (<https://dumps.wikimedia.org/idwiki/20240820/>) mempunyai saiz sebesar 1008.3 MB., iaitu tiga kali lebih besar daripada Bahasa Melayu. Sementara Wikipedia *English* pula mempunyai saiz sebesar 22.5 GB. Demikian jugalah bahasa-bahasa utama yang lain.

Secara keseluruhan nya data-data daripada Wikipedia masih kecil untuk kegunaan pembelajaran mesin di dalam LLM. Sementara itu, data daripada Wikipedia Bahasa Melayu, walaupun lebih “bersih” daripada aspek bahasa, mempunyai dua permasalahan:

- a) Ia disunting secara terbuka oleh sesiapa sahaja. Ini bermakna tiada piawaian yang dibuat untuk memastikan ketepatan serta keabsahan penggunaan bahasa Melayu.
- b) Untuk permodelan LLM, ia dikategorikan sebagai “*medium resource language*”, iaitu jumlah datanya tidak cukup besar secara tersendiri untuk dibuat sebagai sumber data utama kepada sesuatu model.

9.2 Ketersediaan Korpus Bahasa Melayu Sebagai Sumber Data Pembelajaran LLM

Kami telah membuat kajian ringkas untuk mengkaji apakah sumber-sumber data yang sedia ada untuk membuat pembelajaran LLM Bahasa Melayu yang rasmi. Berikut adalah beberapa rumusan yang kami perolehi.

Korpus DBP: Mengikut paparan daripada halaman web Korpus DBP, berikut adalah perincian yang dipaparkan. Saiz mutakhir (sehingga 25 November 2008) Pangkalan Data Korpus adalah lebih kurang 135 juta perkataan yang terkandung dalam sepuluh subkorpus seperti yang berikut:

Bil	Subkorpus	Jumlah Semasa	Jenis Bahan
1	Buku	31,580,305	Novel, buku ilmiah, buku umum, buku teks
2	Majalah	14,406,888	Umum meliputi pelbagai bidang
3	Akhbar	80,029,347	Harian, tabloid, mingguan
4	Terjemahan (buku)	2,021,191	Buku ilmiah, buku umum
5	Efemeral	290,207	Risalah, brosur, iklan
6	Drama	404,176	Drama yang dibukukan
7	Puisi	116,428	Puisi yang dibukukan
8	Kad Bahan	3,130,641	Kad kutipan untuk penyusunan Kamus Dewan
9	Teks Tradisional	2,825,329	Teks lama dalam bentuk hikayat, cerita rakyat
10	Buku Teks	1,095,726	Buku teks peringkat rendah dan menengah

Sumber: <http://rujukan.dbp.gov.my/>

Kamus dan Tesaurus Bahasa Melayu: Berikut adalah beberapa dapatan daripada hasil carian yang dibuat mengenai Kamus dan Tesaurus Bahasa Melayu yang tersedia ada.

→ Kamus Dewan Edisi Keempat (DBP, 2007)

Edisi terkini Kamus Dewan memaparkan sekitar 82,900 kata masukan (34,578 entri, 24,180 subentri, dan 24,142 frasa entri), 5,843 kata masukan baru yang berupa: kata umum, istilah umum pelbagai bidang dan ilmu, kata serapan Arab, kata slanga, dan 375 kata yang mengalami peluasan makna. Ia juga menjalankan pengemaskinian ejaan serta pembaikan takrif, bersesuaian dengan perubahan dan perkembangan yang dialami oleh bahasa Melayu.

→ **Tesaurus Bahasa Melayu Dewan Edisi Baharu (DBP, 2015)**

Tesaurus ini merupakan tesaurus Bahasa Melayu yang komprehensif yang mengandungi sekitar 5400 entri utama dan 18000 subentri. Setiap entri dan subentri dilengkapi dengan pilihan sinonim dan kata-kata yang berkaitan, dengan lebih tersusun dan jelas berdasarkan pecahan makna dan kolokasi (mengikut konteks). Tesaurus ini juga dilengkapi dengan kelas kata dan label geografi.

→ **Kamus Dewan Perdana (DBP, 2020)**

Menurut Kamus Dewan Perdana (DBP, 2020) jumlah kesemua perkataan dalam bahasa Melayu melebihi 120,000 entri dan subentri dan jumlah tersebut bertambah dari masa ke masa mengikut perkembangan penggunaan bahasa tersebut oleh masyarakat penuturnya. Entri dan subentri ini termasuklah kata umum, kata nama khas, istilah, kata dialek, kata pinjaman, kata klasik dan sebagainya. Jumlah 120,000 entri/subentri yang berpandukan Sistem Ejaan Rumi, Pedoman Umum Ejaan Jawi dan Pedoman Umum Sebutan Baku Bahasa Melayu yang terkandung di dalam Daftar Kata Bahasa Melayu, Rumi-Sebutan-Jawi Edisi Kedua (2008, 2016), dan memanfaatkan data korpus dalam penyusunan entrinya.

→ **Set Data Malaya oleh Mesolitica:**

Satu usaha yang telah dilaksanakan oleh Mesolitica untuk mengumpul data untuk kegunaan NLP dan LLM Bahasa Melayu di bawah: <https://github.com/mesolitica/malaysian-dataset>. Data-data tersebut telah juga diletakkan di bawah platform sumber terbuka Hugging Face (<https://huggingface.co/mesolitica>) dimana terdapat (pada masa terkini) sebanyak 197 set data yang telah dipersiapkan hasil daripada pelbagai tugas.

Mesolitica juga telah membuat pemprosesan data berbentuk teks sebagai bahan pembelajaran LLM dan telah dikumpulkan dibawah repositori Hugging Face sebagai: <https://huggingface.co/datasets/malaysiaai/pretrain-text-dataset/tree/main>

Set data tersebut dikumpul melalui pelbagai web crawl daripada halaman web daripada Malaysia. Data-data yang dikumpul kebanyakan daripada berbagai bahasa, terutamanya Bahasa Melayu, Bahasa English, dan juga Bahasa Tamil, Hokkien, serta bahasa kegunaan didalam media sosial seperti Tweeter dan lain-lain.

Pada hemat kami, setakat ini pengumpulan data yang dibuat oleh mereka adalah yang terbesar untuk Bahasa Melayu dan juga di Malaysia.

- **Projek “Malay Concordance”:** Projek Malay Concordance telah diusahakan oleh Australian National University (ANU) (Rujukan: <https://mcp.anu.edu.au/>). Keunikan projek MCP ini adalah dimana teks-teks yang dikumpulkan adalah daripada dokumen-dokumen bersejarah daripada period 1250 sehinggalah tahun 2000. Naskah-naskah yang ada telahpun didigitalkan. Keseluruhan teks mempunyai 5,889,508 perkataan, 140,000 ayat, daripada 553,427 dokumen, 648,651 keratan akhbar, dan lain-lain.
- **Projek “WordNet Bahasa”:** Projek Wordnet Bahasa telah dibangunkan oleh Nanyang Technological University, Singapura (Rujukan: <https://wn-msa.sourceforge.net/index.eng.html>). Ia merupakan gabungan Bahasa Melayu dan Indonesia. Projek ini merupakan satu cabang daripada projek “Princeton WordNet”. Malangnya projek ini tidak diteruskan lagi pada masakini. Walaubagaimana pun data-data boleh dimuat turun daripada halaman web mereka.

WordNet Bahasa mempunyai 49,668 synsets, 145,696 senses, dan 64,431 perkataan yang unik.

10. Ujian Terhadap LLM Sedia Ada Mengenai Bahasa Melayu

Di dalam seksyen ini, kami membawa beberapa contoh ujian Bahasa Melayu dengan menggunakan beberapa LLM terpilih. Secara umumnya, ujian bahasa kepada LLM adalah di dalam kemampuan LLM tersebut memberikan penjelasan terhadap maksud kepada sesuatu perkataan dalam sesuatu bahasa – sama ada ia bertepatan dengan maksud yang rasmi seperti yang terdapat di dalam kamus bahasa tersebut atau tidak.

Kedua, kami akan menguji terjemahan satu teks dalam Bahasa Indonesia yang akan diterjemahkan kepada Bahasa Melayu menggunakan LLM tersebut. Ketiga, kami akan menguji keupayaan LLM tersebut dalam membuat ringkasan berdasarkan terjemahan yang dibuat sebelumnya.

Untuk pilihan model LLM, kami memilih model-model berikut:

Nama Model	Sumber	Nota
ChatGPT	Melalui API berbayar	OpenAI Model GPT4o
Llama 3.1	Meta AI	Llama3.1:8B
Gemma 2	Google	Gemma2:27B
Mistral	Mistral AI	Mistral:7B
Phi3-medium	Microsoft	Phi3-m:7.9B

10.1 Ujian Maksud Perkataan

Menguji maksud perkataan: **Topang**

Definisi daripada Kamus Dewan Edisi Keempat adalah:

Sangga (penyokong) yg bercabang; tiang ~ tiang yg di hadapan; bertopang 1. memakai topang (sangga), disangga, disokong, ditunjang: dahan pokok itu harus diberi ~ kalau tidak nanti patah; (duduk) ~ dagu (duduk) dgn dagunya disangga dgn tangan; 2. bertumpu (pd), berdasar (pd), bertekan (pd): tangannya ~ pd kerusi; ~ lutut duduk dgn meletakkan tangan (siku) pd lutut; bertopangkan menggunakan sesuatu sbg topang; menopang menyangga (menunjang) dgn kayu dll yg bercabang: bahagian hadapan rumah itu ditopang dgn dua batang kayu; tangannya ~ kepalanya; Raja Shahrul Alam ~ dagu seraya mengurut-urut janggutnya; topang-~ bantu-membantu, sokong-menyokong, tolong-menolong; topangan penopang; penopang sokong, penunjang, penyangga, penyokong.

Sementara definisi mengikut Wikikamus Bahasa Indonesia adalah:

Penyokong atau penunjang (kayu dsb.) yang bercagak: Karena buahnya sarat, pokok pisang itu terpaksa ditunjang dengan topang

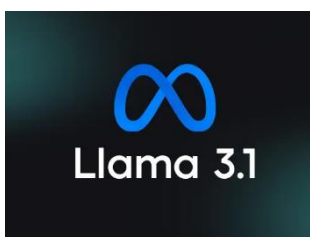
Soalan» Maksud topang dalam Bahasa Melayu.

Jawapan ChatGPT:



Dalam Bahasa Melayu, “topang” merujuk kepada alat atau objek yang digunakan untuk menyokong atau menahan sesuatu agar tetap berdiri atau berada dalam kedudukan tertentu. Contoh kegunaan “topang” adalah seperti topang untuk menyokong pokok yang baru ditanam, atau topang untuk membantu seseorang yang cedera agar dapat berjalan. Dalam erti kata yang lebih kiasan, “topang” juga boleh merujuk kepada sokongan atau bantuan yang diberikan kepada seseorang atau sesuatu.

Jawapan Llama 3.1:



Dalam bahasa Melayu, “topang” memiliki beberapa makna bergantung pada konteksnya. Berikut beberapa kemungkinan:

1. **Bantu:** Istilah “topang” sering digunakan untuk menggambarkan perbuatan membantu atau memberi sokongan kepada seseorang atau sesuatu. Contoh: “Saya akan topang kamu dalam projek ini.”
2. **Mengalihkan perhatian:** Dalam beberapa konteks, “topang” juga dapat berarti mengalihkan perhatian dari suatu hal yang tidak diinginkan kepada sesuatu yang lain. Contoh: “Dia topang perhatiannya dengan membicarakan tentang hujan untuk mengalihkan perhatian daripada masalah itu.”
3. **Mengelirukan atau menyesatkan:** Istilah ini juga dapat digunakan dalam konteks yang lebih negatif, seperti mengelirukan atau menyesatkan. Contoh: “Dia membicarakan topang untuk mengelirukan saya.”

Namun, perlu diingat bahawa makna terakhir ini lebih bersifat negatif dan tidak selalu diterima dengan baik dalam percakapan sehari-hari.

Dalam konteks yang lebih umum, seperti dalam kalimat “Topang kami!”, “topang” digunakan sebagai ungkapan untuk mengajak semua orang bekerja sama atau membantu sesuatu.

Jawapan Gemma 2:



“Topang” dalam Bahasa Melayu bermakna **menopang**, **mendukung**, atau **memberi kekuatan**.

Ia boleh digunakan dalam beberapa konteks, seperti:

- **Menopang fizikal:** “Dia menggunakan tongkat untuk **menopang** dirinya ketika berjalan.”
- **Mendukung idea:** “Argumennya **ditopang** oleh bukti yang kukuh.”
- **Memberi kekuatan:** “Cinta dan kasih sayang ibu **menopang** semangat anaknya.”

Jawapan Mistral:

Mungkin anda tertanya-tanya tentang apa itu “Makna Topang” dalam Bahasa Melayu. Namun, dalam konteks yang umum, kalau kita menggunakan perkataan “Topang”, ia merujuk kepada sebuah alat peledak atau senjata api jenis revolver. Jadi, “Makna Topang” dalam Bahasa Melayu berarti “Artikel atau penjelasan mengenai senjata api jenis revolver.”

Ini adalah contoh bertulis:

- Makna Topang ialah alat peledak yang memiliki banyak hujung peledak yang dapat diputar sekali.
- Senjata api jenis topang merupakan senjata pelindung diri yang sangat populer dan diperlukan oleh kebanyakan negara dunia.

Jawapan Phi 3:

Kata-katanya adal cuaca yang berpendapat. Dalam Bahasa Melayu, kita dapat menyatakan ini sebagai “Maka agar dia dianggap sebagai cuaca yang berdiri”. Contoh pengulangan untuk 50 warna sungai adalah mengikuti sumbu nila ke utara, dan seterusnya.

10.2 Terjemahan

Teks berikut adalah daripada Bahasa Indonesia (sumber: https://mylanguages.org/indonesian_reading.php)

Setiap orang berhak mendapat pendidikan. Pendidikan harus gratis, setidaknya-tidaknya untuk tingkat sekolah rendah dan pendidikan dasar. Pendidikan rendah harus diwajibkan. Pendidikan teknik dan jurusan secara umum harus terbuka bagi semua orang, dan pengajaran tinggi harus secara adil dapat diakses oleh semua orang, berdasarkan kepantasan. Pendidikan harus ditujukan ke arah perkembangan pribadi yang seluasluasnya serta memperkokoh rasa penghargaan terhadap hak-hak manusia dan kebebasan asasi. Pendidikan harus menggalakkan saling pengertian, toleransi dan persahabatan di antara semua bangsa, kelompok ras maupun agama, serta harus memajukan kegiatan Perserikatan Bangsa-Bangsa dalam memelihara perdamaian. Orang-tua mempunyai hak utama untuk memilih jenis pendidikan yang akan diberikan kepada anak-anak mereka. Setiap orang berhak untuk turut serta dengan bebas dalam kehidupan kebudayaan masyarakat, untuk mengecap kenikmatan kesenian dan berbagi dalam kemajuan ilmu pengetahuan dan manfaatnya. Setiap orang berhak untuk memperoleh perlindungan atas kepentingan-kepentingan moril dan material yang diperoleh sebagai hasil dari sesuatu produksi ilmiah, kesusasteraan atau kesenian yang diciptakannya.

Kita akan meminta LLM untuk: **Terjemahkan berikut ke dalam Bahasa Melayu “teks di atas”.**

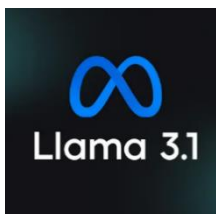
**Terjemahan
ChatGPT:**



Berikut adalah terjemahan teks tersebut ke dalam Bahasa Melayu:

“Setiap orang berhak mendapat pendidikan. Pendidikan harus percuma, sekurang-kurangnya untuk per-ingkat sekolah rendah dan pendidikan asas. Pendidikan rendah harus diwajibkan. Pendidikan teknikal dan vokasional secara umumnya harus terbuka kepada semua orang, dan pengajian tinggi harus dapat diakses dengan adil oleh semua orang, berdasarkan kelayakan. Pendidikan harus ditujukan kepada perkembangan peribadi yang seluas-luasnya serta memperkukuhkan rasa penghargaan terhadap hak-hak manusia dan kebebasan asasi. Pendidikan harus menggalakkan saling pengertian, toleransi, dan persahabatan di antara semua bangsa, kumpulan etnik mahupun agama, serta harus memajukan kegiatan Pertubuhan Bangsa-Bangsa Bersatu dalam memelihara keamanan. Ibu bapa mempunyai hak utama untuk memilih jenis Pendidikan yang akan diberikan kepada anak-anak mereka. Setiap orang berhak untuk turut serta dengan bebas dalam kehidupan budaya masyarakat, untuk menikmati kesenian dan berkongsi dalam kemajuan ilmu pengetahuan dan manfaatnya. Setiap orang berhak untuk memperoleh perlindungan atas kepentingan moral dan material yang diperoleh sebagai hasil daripada sesuatu pengeluaran ilmiah, kesusasteraan, atau kesenian yang diciptakannya.”

**Terjemahan Llama
3.1:**



“Setiap orang berhak mendapat pendidikan. Pendidikan haruslah percuma, sekurang-kurangnya untuk tahap sekolah rendah dan pendidikan dasar. Pendidikan rendah harus diwajibkan. Pendidikan teknik dan jurusan secara umum harus dibuka bagi semua orang, dan pengajaran tinggi harus dapat diakses oleh semua orang dengan adil, berdasarkan kelayakan. Pendidikan harus ditujukan ke arah perkembangan peribadi yang luas serta memperkokoh rasa menghargai hak-hak manusia dan kebebasan asasi. Pendidikan harus menggalakkan saling pengertian, toleransi dan persahabatan di antara semua bangsa, kumpulan kaum atau agama, serta harus memajukan kegiatan Pertubuhan Bangsa-Bangsa dalam menjaga perdamaian. Orang tua mempunyai hak utama untuk memilih jenis pendidikan yang akan diberikan kepada anak-anak mereka. Setiap orang berhak untuk turut serta dengan bebas dalam kehidupan kebudayaan masyarakat, untuk menikmati kegembiraan kesenian dan berbagi dalam kemajuan ilmu pengetahuan dan manfaatnya. Setiap orang berhak untuk memperoleh perlindungan atas kepentingan-kepentingan moril dan materiel yang diperoleh sebagai hasil dari sesuatu produksi ilmiah, kesusasteraan atau kesenian yang diciptakan oleh mereka.”

Terjemahan Gemma 2:



“Setiap insan berhak mendapat pendidikan. Pendidikan mestilah percuma, sekurang-kurangnya untuk peringkat sekolah rendah dan pendidikan asas. Pendidikan rendah perlu diwajibkan. Pendidikan teknik dan jurusan secara umum mestilah terbuka kepada semua orang, dan pengajian tinggi mestilah adil dapat diakses oleh semua orang, berdasarkan kebolehan.

Pendidikan hendaklah ditujukan ke arah perkembangan diri yang seluas-luasnya serta memperkukuh rasa penghargaan terhadap hak asasi manusia dan kebebasan asas. Pendidikan hendaklah menggalakkan saling memahami, toleransi dan persahabatan antara semua bangsa, kaum dan agama, serta hendaklah memajukan kegiatan

Pertubuhan Bangsa-Bangsa Bersatu dalam memelihara keamanan. Ibu bapa mempunyai hak utama untuk memilih jenis pendidikan yang akan diberikan kepada anak-anak mereka.

Setiap insan berhak untuk mengambil bahagian secara bebas dalam kehidupan kebudayaan masyarakat, untuk menikmati kesenian dan berkongsi dalam kemajuan ilmu pengetahuan serta manfaatnya. Setiap insan berhak untuk memperoleh perlindungan terhadap kepentingan moral dan material yang diperoleh sebagai hasil daripada sesuatu penciptaan ilmiah, kesusasteraan atau kesenian.”

Terjemahan Mistral:



Mistral menterjemah terlebih dahulu kepada English, kemudian daripada English diterjemahkan pula kepada Bahasa Melayu. Hasil nya adalah seperti berikut:

“Setiap individu mempunyai hak kepada mendapat pendidikan. Pendidikan harus bebas, setidak-tidaknya untuk tingkat sekolah rendah dan pendidikan dasar. Pendidikan rendah harus diwajibkan. Pendidikan teknik dan jurusan secara umum harus terbuka kepada semua orang, dan pengajaran tinggi harus diterima secara adil oleh semua orang berdasarkan kepantasan. Pendidikan harus ditujukan ke arah perkembangan peribadi yang seluas-luasnya serta memperkokoh rasa penghargaan terhadap hak-hak manusia dan kebebasan asasi. Pendidikan harus menggalakkan saling pengertian, toleransi dan persahabatan di antara semua bangsa, kelompok ras maupun agama, serta harus memajukan kegiatan Perserikatan Bangsa-Bangsa dalam memelihara perdamaian. Orang-tua mempunyai hak utama untuk memilih jenis pendidikan yang akan diberikan kepada anak-anak mereka. Setiap individu mempunyai hak kepada turut serta dengan bebas dalam kehidupan kebudayaan masyarakat, untuk mengecap kenikmatan kesenian dan berbagi dalam kemajuan ilmu pengetahuan dan manfaatnya. Setiap individu mempunyai hak kepada perlindungan atas kepentingan-kepentingan moral dan material yang diperoleh sebagai hasil dari sesuatu produksi ilmiah, kesusasteraan atau kesenian yang diciptakannya.”

Terjemahan Phi 3:



“Kata-katanya semua orang mempunyai hak untuk mendapat pendidikan. Pendidikan harus diberi gratis, tetapi setidaknya untuk kelas dasar dan sekolah rendah. Pendidikan sekolah menengah harus ditujukan. Pendidikan teknik dan jurusan umum harus terbuka bagi semua orang, dan pendidikan tinggi harus adil diakses oleh semua orang, dengan mengenal kepantasan yang diperlukan. Pendidikan harus menjadi suatu bentuk perkembangan pribadi untuk membawa sama-sama rasa hormat terhadap hak asasi manusia dan kehakiman. Pendidikan harus membolehkan saling pengertian, toleransi dan persahabatan di antara semua bangsa, rakyat serta agama yang berlawanan. Pendidikan harus mendorong kegiatan Perserikatan Bangsa-Bangsa dalam memelihara perdamaian. Ibu bapa mempunyai hak utama untuk pilih jenis pendidikan yang akan diberi kepada dirinya daripada anak-anaknya. Setiap orang boleh serta dengan bebas dalam kehidupan budaya masyarakat, mengikmu kesenian dan berbagi dalam kemajuan ilmu pengetahuan dan manfaatnya. Setiap orang mempunyai hak untuk mendapat perlindungan atas kepentingan-kepentingan moral, fizikal yang diperoleh sebagai hasil dari pembuatan ilmiah, penulisan dan seni.”

10.3 Ringkasan Atau Kesimpulan

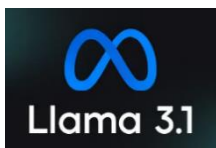
Seterusnya kita akan memberikan arahan untuk **membuat ringkasan kepada teks yang diterjemahkan sebelumnya** sebagai arahan berikutnya.

Ringkasan ChatGPT:



“Setiap orang berhak mendapat pendidikan yang percuma, sekurang-kurangnya pada peringkat sekolah rendah, yang harus diwajibkan. Pendidikan teknikal dan pengajian tinggi perlu dibuka kepada semua berdasarkan kelayakan, dengan tujuan memperkembangkan peribadi, menghargai hak asasi manusia, dan menggalakkan toleransi serta persahabatan antara bangsa dan agama. Ibu bapa berhak memilih jenis pendidikan untuk anak-anak mereka. Selain itu, setiap individu berhak menikmati kebudayaan, kesenian, dan kemajuan ilmu pengetahuan, serta perlindungan ke atas hasil karya ilmiah, kesusasteraan, atau kesenian yang dihasilkan.”

Ringkasan Llama 3.1:



“Setiap orang berhak mendapat pendidikan yang berkualiti dan percuma, sekurang-kurangnya untuk tahap sekolah rendah dan pendidikan dasar. Pendidikan haruslah diwajibkan bagi semua dan dibuka bagi jurusan-jurusan teknis dan tinggi, agar setiap orang dapat mengakses ilmu pengetahuan dengan adil dan bebas.”

Ringkasan Gemma 2:



“Mentari pagi menyapa lembut wajah kota, membangunkanku dari tidur nyenyak. Aroma kopi yang menguar dari dapur membuat perutku keroncongan. Aku bangkit dari tempat tidur dan berjalan menuju jendela, menikmati udara segar dan kicauan burung-burung di pohon rindang di luar rumah. Hari ini terasa penuh harapan, seperti lembaran buku kosong yang menanti untuk diisi dengan cerita indah.”

Terjemahan Mistral:



“Pendidikan harus membina pemimpin-pemimpin yang bertanggung jawab atas kebersihan dan keselamatan alam sekitar. Orang-orang mestilah berkepentingan bagi alam semesta dan menjaga lingkungan ini selamat bagi generasi makhluk hidup yang akan datang.”

Ringkasan Phi 3:



“Pendidikan merupakan pentadbir yang sangat penting dalam kehidupan seseorang. Ia dapat membantu individu untuk mengekalkan pandangan positif terhadap hidup dan mempercepat pergerakan kemajuan masyarakat. Pendidikan juga boleh membantu orang mengambil keputusan yang baik dalam kehidupan seharian, serta memberi anda kekuatan untuk menyiapkan diri sendiri pada masa depan. Selain itu, pendidikan juga penting kerana ia membantu mereka menggali pengetahuan dan memahami dunia yang lebih luas serta memberi anda pelajaran untuk menjalankan pekerjaan yang relevan dengan kecemasan diri.”

11. Usaha-Usaha Yang Sedia Ada Untuk Membina LLM Bahasa Melayu

Kami telah membuat kajian ringkas terhadap beberapa usaha untuk membuat atau membina LLM Bahasa Melayu yang telah dibangunkan sebagai sumber terbuka, seperti melalui platform Github dan juga Hugging Face. Berikut adalah ringkasan yang kami buat mengenai usaha-usaha tersebut.

11.1 Projek Malaya oleh Mesolitica

Projek Malaya merupakan satu repositori (perpustakaan digital) yang berfokuskan Bahasa Melayu terbesar yang telah dibina sebagai sumber terbuka. Berikut adalah jadual ringkas kerja-kerja yang telah mereka sempurnakan (ataupun dalam pembinaan):

Kerja Utama	Nota Penjelasan	Tempat Simpanan	Komentar
Membina set data untuk latihan pembelajaran mesin	Set data yang telah dibersihkan, dikumpulkan dalam format json dari pelbagai sumber	Hugging Face	Kepelbagaian sumber dan proses membersihkan data telah dibuat
Membina modul “ <i>data augmentation</i> ”	Membina modul “ <i>abstractive</i> ”, “ <i>encoders</i> ” serta “ <i>wordvectors</i> ”	Modul Python dan Hugging Face transformers	Belum teruji sepenuhnya
Kamus Bahasa Melayu	Membina sambungan kepada carian umum dalam talian DBP	Modul Python	Hakcipta DBP diperlukan
Modul pembuatan token	Tokenizer perkataan, ayat dan <i>syllable</i>	Modul python	Belum teruji sepenuhnya
Modul LLM	Transformer dan GPT	Hugging Face	Kerja-kerja masih di peringkat awalan
Model pembetulan ejaan	Pelbagai modul dan kaedah	Modul Python	Masih belum diuji sepenuhnya
Modul tulisan Jawi	Model kecil untuk skrip Jawi	Hugging Face	Masih dalam kerja-kerja pembinaan
Modul tatabahasa	Model untuk tagging kesalahan tatabahasa dalam Bahasa Melayu	Modul Python	Masih dalam kerja-kerja pembinaan
Modul pengeluaran teks (<i>text generative</i>)	Salah satu daripada tugas dalam LLM iaitu pembuatan teks	Hugging Face	Kerja-kerja masih dalam pembinaan
Modul pembersihan teks	Pembersihan data dari pelbagai sumber perlu ditapis; seperti sumber Twitter dll	Hugging Face	Sudah hampir sempurna

Kerja Utama	Nota Penjelasan	Tempat Simpanan	Komentar
Modul “ <i>Embeddings</i> ”	Proses pembuatan “ <i>word vectors</i> ”, “ <i>embeddings</i> ” dll di dalam NLP	Modul Python	Kaedah yang diguna pakai masih perlu tambah baik
Modul “ <i>Tagging</i> ”	Proses penamaan entity (NER), POS, dll	Modul Python	Masih dalam kerja-kerja pembaikan
Modul “ <i>Knowledge Graph</i> ”	Pembinaan kaedah membina “ <i>knowledge graph</i> ” daripada teks	Modul Python	Masih di peringkat awalan
Modul terjemahan	Pembinaan terjemahan secara automatik	Modul Python	Masih dalam peringkat awalan

Sumber: <https://malaya.readthedocs.io>

Secara keseluruhan nya, kerja-kerja yang telah dibuat oleh Mesolitica melalui Projek Malaya amatlah membanggakan.

Percubaan yang mereka laksanakan untuk pelbagai kerja-kerja NLP untuk Bahasa Melayu adalah suatu yang perlu didokong. Walaubagaimanapun satu kekurangan yang ketara ialah mereka menggunakan pelbagai teknik dan kaedah secara “*try and error*” tanpa panduan yang jelas dan pembentukan metodologi yang disokong oleh pakar-pakar bahasa Melayu dan juga pakar-pakar AI/ML. Usaha mereka ini perlu digabungkan dalam usaha membina LLM Bahasa Melayu Rasmi yang dicadangkan.

11.2 Projek MelayuBERT

Projek MelayuBERT diusahakan oleh Steven Limcorn daripada University of Queensland, Australia. Data pembelajaran diperolehi daripada Open Super-large Crawled ALMAnaCH coRpus (OSCAR) menggunakan struktur “*goclassy*”. Model yang digunakan ialah English BERT, disunting halus (*fine-tune*) menggunakan set data daripada Bahasa Melayu. Pembinaan model telah berhenti setakat ini. Ia disimpan di Hugging Face (<https://huggingface.co/StevenLimcorn/MelayuBERT>).

12. Kenapa Perlu Membina LLM Bahasa Melayu Rasmi?

Permasalahan yang ketara terhadap LLM yang sedia ada adalah kecenderungan kepada Bahasa Inggeris (*English biased*) dan juga kepada bahasa negara barat (*western biased*). Ini adalah disebabkan peratusan data latihan dan pembelajaran mesin lebih daripada 70 peratus adalah daripada bahasa-bahasa utama negara barat, terutamanya bahasa English. Untuk bahasa-bahasa Asia Tenggara pula, Bahasa Melayu dan Bahasa Indonesia bercampur baur diantara keduanya menyebabkan kecenderungan adalah memihak kepada Bahasa Indonesia (*Indonesian biased*).

Dalam jangka waktu kedepan, perkara-perkara ini akan lebih memuncak dan bakal menyebabkan proses penghanyutan (*drifting*) berlaku dalam kalangan pengguna. Terutamanya apabila penggunaan LLM yang sudah dipasarkan seperti ChatGPT dan lain-lain yang sudah menular dalam kalangan pengguna natif Bahasa Melayu. Survei daripada Stanford Human AI Institute menunjukkan sudah lebih 60 peratus daripada pengguna di Malaysia menggunakan ChatGPT atau seumpamanya.

Sejak kebelakangan ini, negara-negara besar seperti negara di Eropah melalui Kesatuan Eropah (EU) sudah mula merangka undang-undang baru untuk mengawal selia kecerdasan buatan, melalui pelbagai inisiatif seperti “Responsible AI”, pembentukan “*guard rails*” dan sebagainya. Demikian juga pengamal bahasa seperti Bahasa German, demikian juga bahasa-bahasa negara Nordik, kesemuanya sudah mengambil langkah untuk memastikan bahasa natif mereka terkawal dalam dunia digital melalui cabaran penaklukan (*invasion*) melalui Generative AI dan LLM seperti ChatGPT dan sebagainya.

Negara China juga telahpun melancarkan LLM mereka sendiri di bawah projek 01.ai, sebagai usaha membina dan memperkenalkan LLM berpiawai Bahasa Mandarin yang dibina oleh mereka sendiri (lihat: <https://www.01.ai/> dan <https://huggingface.co/01-ai>). Usaha ini dipelopori oleh Kai-fu Lee, seorang pakar dalam bidang kecerdasan pintar negara China.

Perkembangan LLM dan GAI terus berkembang pesat dengan pelbagai aplikasi yang diperkenalkan kepada bidang-bidang (*domains*) yang lebih khusus seperti dalam bidang pendidikan, kesihatan, perubatan, dan lain-lain. Pada pandangan kami, penggunaan yang lebih spesifik akan menambahkan lagi kerumitan kepada kedudukan serta fungsi Bahasa Melayu didalam dunia digital dan juga dalam perkembangan kecerdasan buatan pintar yang semakin menyemarak pada ketika ini dan masa depan.

Memandangkan pada masa kini, tiada suatu pun usaha yang giat serta fokus kepada permasalahan ini diberikan perhatian dan tumpuan khusus oleh mana-mana badan, organisasi, atau institusi dalam negara. Perkara ini perlu ditangani dengan segera sebagai satu usaha yang kritikal dan perlu dimajukan dengan segera.