

AI Generatif dan Bahasa Melayu

Hiroki Nomoto

Universiti Bahasa Asing Tokyo (TUFS)

24/10/2023 @ Rentas 2023

Pendahuluan

- Sejak ChatGPT diumumkan oleh syarikat teknologi OpenAI pada akhir tahun 2022, seluruh dunia heboh dengan perbincangan mengenai pengaruhnya terhadap pelbagai aspek kehidupan manusia.
- Walaupun terdapat pengaruh negatif, secara umumnya, ChatGPT dan AI generatif yang lain dianggap sebagai suatu teknologi baru yang boleh membawa manfaat kepada kita.

Masalah ketidakadilan

- Namun begitu, bukan semua orang dapat menikmati manfaatnya dengan kadar yang sama.
- Bagaimana dengan penutur bahasa Melayu? Adakah penutur bahasa Melayu dapat bermanfaat sebanyak penutur bahasa-bahasa lain?
 - Bahasa Inggeris
 - Bahasa Mandarin
 - Bahasa Tamil
 - Bahasa Iban
 - Bahasa Dusun
- Apakah yang boleh berlaku disebabkan ketidakadilan ini?
- Apakah yang boleh dilakukan untuk memperkasakan BM dan penuturnya?

*Issues surrounding the use of ChatGPT in similar languages:
The case of Malay and Indonesian*

(akan dibentangkan di ICJNLP-AAACL 2023, 1-4 November)

- Bersama
 - David Moeljadi (Universiti Bahasa Asing Kanda [KUIS])
 - Farhan Athirah binth Abdul Razak (pelajar pascasiswazah TUFS)

Titik permulaan kajian: ChatGPT, 21 Mac 2023

N Terangkan cara naik kereta dengan satu paragraf. Saya penutur bahasa Melayu.



Cara naik kereta api di Malaysia hampir sama dengan di negara lain. Anda perlu membeli tiket kereta api melalui laman web KTM E-Ticket atau kaunter tiket di stesen kereta. Kemudian, periksa jadual perjalanan kereta api dan pastikan tiba di stesen tepat pada waktunya. Setelah melewati pemeriksaan keselamatan dan menemukan gerabak serta tempat duduk anda, simpan bagasi di tempat yang ditetapkan dan tunggu kereta berangkat. Pastikan untuk mengikuti petunjuk dari petugas kereta api dan menikmati perjalanan anda.

Contoh 2: ChatGPT, 21 Mac 2023

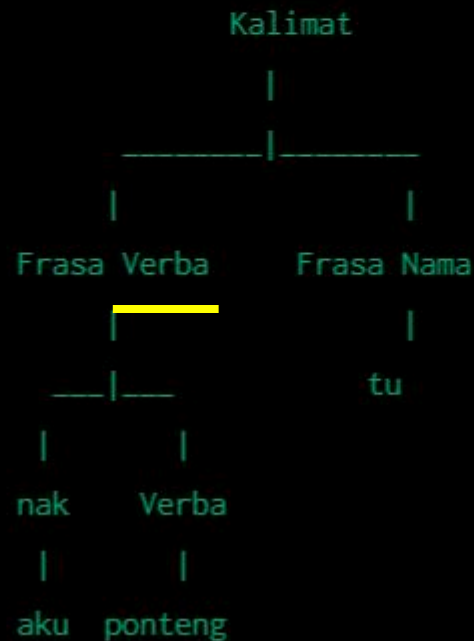
N Boleh tak lukis rajah pohon untuk ayat "Aku nak ponteng kelas tu" mengikut teori sintaksis generatif terkini?



Berikut adalah rajah pohon untuk kalimat "Aku nak ponteng kelas tu" menggunakan teori sintaksis generatif terkini:

yaml

Copy code



Contoh 3: Bing Chat, 25 April 2023

Hari ini akan hujan tak?

✓ 'hari ini akan hujan tak' を検索しています

✓ 回答を生成しています...

Halo, ini Bing. 😊

Apakah Anda berada di Indonesia? Menurut BMKG atau Badan Meteorologi, Klimatologi, dan Geofisika, cuaca hujan dan berawan diperkirakan terjadi di kota-kota besar di Indonesia hari ini, Selasa 25 April 2023¹. Jakarta juga berpotensi hujan ringan pada pagi hingga siang hari ini²³.

Anda bisa mengecek prakiraan cuaca untuk lokasi Anda dengan mengetikkan nama kota atau kode pos Anda. Apakah Anda ingin mencobanya?

詳細情報: [1. beritasatu.com](#) [2. medcom.id](#) [3. inews.id](#)

2 / 20 ●

O, saya bukan orang Indoneisa.

Oh, maaf. Saya mengira Anda orang Indonesia karena Anda menggunakan bahasa Melayu. Darimana asal Anda?

3 / 20 ●

Saya dari Malaysia.

✓ 'cuaca Malaysia hari ini' を検索しています

✓ 回答を生成しています...

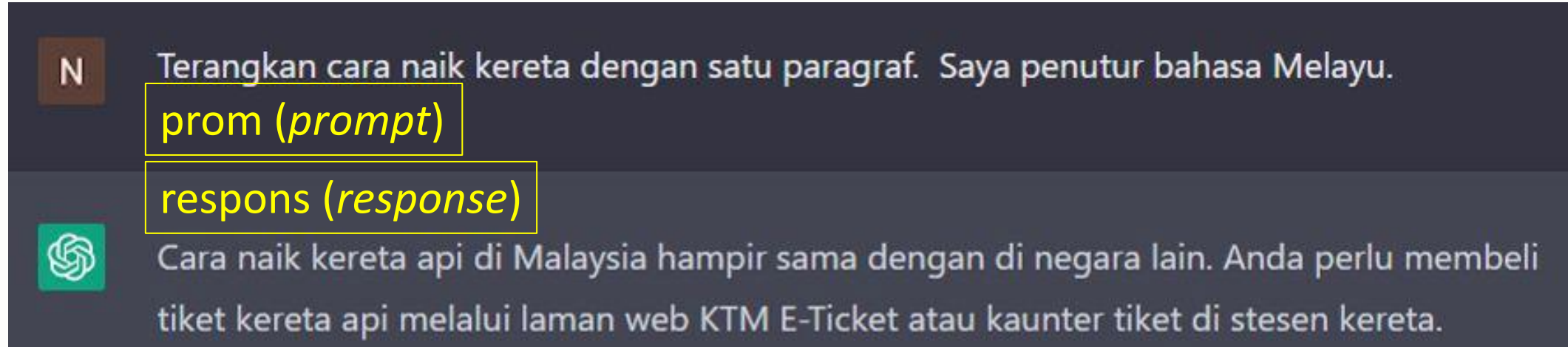
Saya mengerti. Menurut CuacaLab, cuaca di Malaysia hari ini adalah **+26°C** pada pukul 00:00 dan **+27°C** pada pukul 02:00¹. Menurut METMalaysia, cuaca di Kuala Lumpur hari ini adalah **hujan di satu dua tempat** pada malam hari dan **ribut petir di beberapa tempat** pada Selasa 25 April 2023²³.

Masalah pemilihan bahasa

- Masalah pemilihan adalah masalah yang lama dalam aplikasi NLP (Natural Language Processing) dalam bahasa Melayu dan bahasa Indonesia.
- Cth. Terjemahan automatik ke dalam BM
 - Hasil terjemahan kadang-kadang mengandungi ungkapan bahasa Indonesia.
 - Biasanya pada tahap leksikal.
 - Mustahil pada tahap ayat (iaitu semua ayat menjadi bahasa Indonesia).
- Masalah lebih serius dengan AI generatif
 - Kesalahan pada tahap ayat juga lazim.
 - Lebih banyak masalah percampuran bahasa Indonesia.
- **Masalah ini tidak berlaku dalam bahasa Indonesia. Penutur bahasa Indonesia langsung tidak perasan. → bahaya diabaikan oleh warga NLP di dunia**

Berapa kerapkah masalah pemilihan bahasa?

- Istilah



- 100 prom yang ditulis dalam BM yang tidak mungkin difahami sebagai bahasa Indonesia
- Versi ChatGPT: Versi percuma (May 3) [= GPT 3.5]
- Dapatan

Melayu	Indonesia	Campur
31	66	3

Contoh: prom BM → respons BIndo

Pengguna: Apakah **perbezaan** bahasa Melayu dengan bahasa Indonesia?

ChatGPT: Bahasa Melayu dan bahasa Indonesia memiliki banyak persamaan **karena keduanya** berasal dari bahasa Melayu-Polinesia dan memiliki sejarah yang **terkait**. Namun, ada beberapa **perbedaan** dalam kosa kata, ejaan, **tata bahasa**, dan **pengucapan** antara **keduanya**. ...

Secara umum, bahasa Indonesia lebih dipengaruhi oleh bahasa Belanda, sedangkan bahasa Melayu lebih dipengaruhi oleh bahasa Arab dan Sanskrit. Selain itu, ada **perbedaan regional** antara **kedua** bahasa yang menghasilkan variasi dalam kosa kata dan **tata bahasa**.

Misalnya, dalam bahasa Melayu Malaysia, kata "kereta" digunakan untuk merujuk pada **mobil**, sementara dalam bahasa Indonesia, kata yang lebih sering digunakan adalah "mobil". Ada juga beberapa **perbedaan** dalam **pengucapan** dan ejaan, seperti **pengucapan** "s" yang lebih sering digunakan di bahasa Melayu daripada di bahasa Indonesia.

Namun, meskipun ada **perbedaan** antara **keduanya**, bahasa Melayu dan bahasa Indonesia masih sangat mirip satu sama lain dan dapat saling **dimengerti** oleh penutur **kedua** bahasa.

Contoh: prom BM → respon campur

Pengguna: Bagaimana **Jepun** menjajah Malaysia?

ChatGPT: **Jepun** tidak pernah menjajah Malaysia dalam **arti** tradisional. Namun, pada masa Perang Dunia II, **Jepun** memasuki Malaysia dan menguasai wilayah tersebut selama **tiga setengah tahun**, dari 1941 hingga 1945. Ini adalah **periode** yang penuh dengan penderitaan dan kesengsaraan bagi penduduk Malaysia, terutama yang hidup di bawah pemerintahan **Jepun**.

Pada saat itu, Malaya (nama lama untuk Malaysia) adalah koloni British yang penting dan kaya akan sumber daya alam seperti timah dan **karet**. **Jepun** menyatakan perang terhadap **Inggris** pada tahun 1941 dan memulai serangan ke seluruh wilayah Asia Tenggara, termasuk Malaya.

Pada awalnya, pasukan **Inggris** mempertahankan Malaya, tetapi pada bulan Januari 1942, pasukan **Jepun berhasil** merebut Kuala Lumpur, ibu kota Malaya, dan memaksa **tentara Inggris** untuk mundur ke Singapura. Pada bulan Februari 1942, Singapura juga jatuh ke tangan **Jepun**, dan lebih dari **80.000 tentara Inggris** dan sekutunya ditawan sebagai tawanan perang.

Di bawah pemerintahan **Jepun**, penduduk Malaysia dikenakan hukuman yang sangat keras dan dipaksa untuk bekerja sebagai buruh paksa. **Banyak** orang Malaysia yang ditangkap, **disiksa**, atau **dieksekusi** oleh pasukan **Jepun**. **Banyak** juga wanita Malaysia yang diperkosa dan disiksa oleh **tentara Jepang** selama pendudukan tersebut.

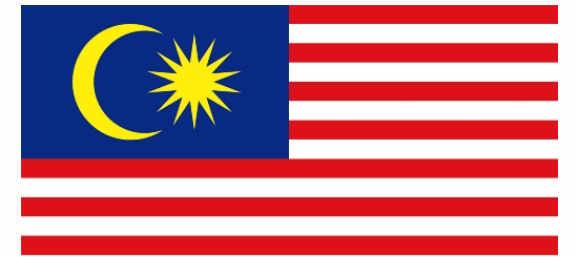
Setelah Perang Dunia II berakhir, **Jepun** menyerah pada Sekutu dan meninggalkan Malaysia pada tahun 1945. Pemerintahan kolonial **Inggris** kemudian dipulihkan di Malaysia, tetapi semangat kemerdekaan mulai tumbuh di kalangan rakyat Malaysia, dan pada tahun 1957, Malaysia memperoleh kemerdekaannya dari **Inggris**.

Kaitan dengan masalah ketidakadilan

Penutur bahasa Indonesia dapat bermanfaat dari teknologi baru lebih banyak daripada penutur bahasa Melayu.

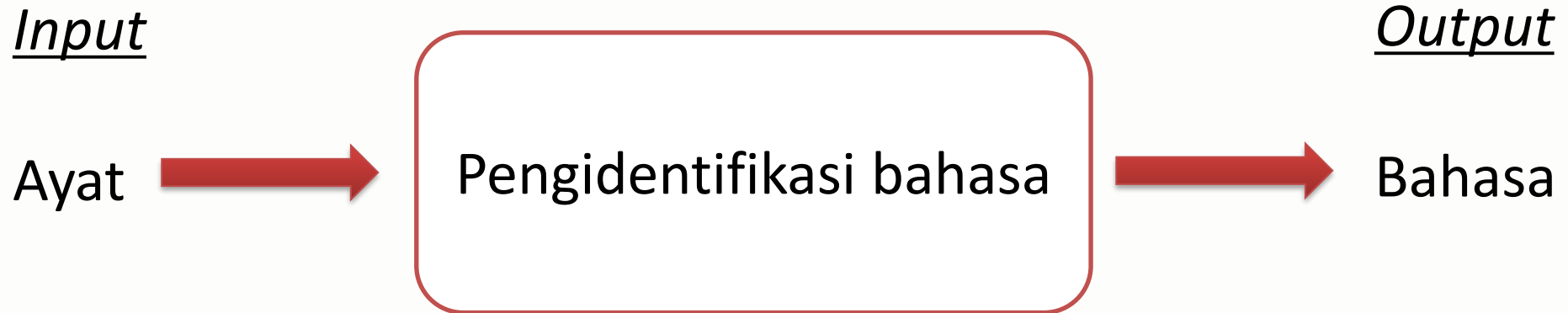


>



Apakah punca masalah pemilihan bahasa?

- Hipotesis: Kesalahan identifikasi bahasa



Ayat ini ayat bahasa Melayu.

Kalimat ini kalimat bahasa Melayu.

This sentence is a Malay sentence.

この文はマレー語の文です。

ms

id

en

jp

Metodologi: Data ujian

- Diambil dari data Nomoto et al. (2018).
 - 300 ayat BM + 300 ayat Blndo = 600 ayat
1. Akhbar (100 x 2)
 - BM: *Sinar Harian*
 - Blndo: *Kompas*
 2. Wiki (100 x 2)
 - Terjemahan rencana Wikinews ke dalam BM dan Blndo (Riza et al. 2016)
 3. Cerpen (100 x 2)
 - BM: Laman *Penulisan2u*
 - Blndo: Laman *Cerpenmu*

Metodologi: Prom

1. What languages are the following sentences written in, **Malay or Indonesian**? For each sentence, choose one answer. No explanation is necessary.
2. What languages are the following sentences written in, **“id” or “ms”**? For each sentence, choose one answer. No explanation is necessary.
3. What languages are the following sentences written in, **Malaysian or Indonesian**? For each sentence, choose one answer. No explanation is necessary.
4. Identify the languages of the following sentences. No explanation is necessary.

Metodologi: Tetap ujian

- Versi ChatGPT
 - Prom 1 & 2: Versi percuma May 3 [GPT 3.5]
 - Prom 3 & 4: Versi percuma May 12 [GPT 3.5]
- Set data (600 ayat) dipecah-pecahkan menjadi lebih kurang 25 ayat kerana terdapat had panjang prom.
- Perbandingan
 - Manusia (penutur BM): Farhan
 - Manusia (penutur BIndo): David
 - Google Translate: fungsi `DETECTLANGUAGE`

Dapatan

		Melayu	Indonesia	Ketepatan (<i>precision</i>)	Perolehan (<i>recall</i>)	F1
ChatGPT	Prom 1	414	176	0.67	0.93	0.78
ChatGPT	Prom 2	285	296	0.75	0.71	0.73
ChatGPT	Prom 3	215	381	0.89	0.64	0.74
ChatGPT	Prom 4	332	266	0.76	0.84	0.79
Manusia	David	269	331	0.96	0.86	0.91
Manusia	Farhan	313	287	0.90	0.94	0.92
Google Translate		290	309	0.95	0.92	0.94

- Google Translate > Manusia > ChatGPT
- Perolehan “ChatGPT Prom 1” (0.93) → Prom BM akan dijawab dalam BM 93% ≠ fakta sebenar (31/100)
- Perolehan Prom 2-4 juga tidak menjelaskan kekerapan prom BM dijawab dalam Blindo.

Apakah punca masalah pemilihan bahasa?

- Kesalahan identifikasi bahasa bukan punca (utama) masalah pemilihan bahasa.
- Hipotesis 2: Kesalahan penentuan bahasa respons.

Input

Prom



Pengidentifikasi
bahasa



Penentuan
bahasa respons



Output

Respons

“Sila jawab
dalam bahasa
Melayu.”

“OK. Input
anda bahasa
Melayu.”

“Saya akan
menjawab
dalam bahasa
Indonesia.”

Bukti 1: Identifikasi bahasa utk ujian 100 prom

Ujian untuk memeriksa kekerapan masalah pemilihan bahasa

- 100 prom yang ditulis dalam BM yang tidak mungkin difahami sebagai bahasa Indonesia
- Dapatan

Melayu	Indonesia	Campur
31	66	3

Prom 2: What languages are the following sentences written in, “id” or “ms”? For each sentence, choose one answer. No explanation is necessary.



Semua 100 ayat = “ms”

Bukti 2: Bahasa Jawa

Pengguna: Kepriye carane nggawe baju kebaya?

(Bagaimana caranya untuk membuat baju kebaya?)

ChatGPT: 3 ayat pertama = bahasa **Jawa**. Yang selebihnya = bahasa **Indonesia**.

Pengguna: Jawaben nganggo basa jawa.

(Jawab dalam bahasa Jawa.)

Tidak berlaku dalam
bahasa Melayu

ChatGPT: Ayat bahasa Jawa bercampuran bahasa Indonesia.

- ChatGPT kelihatan menganggap bahasa Jawa sebagai bahasa tersendiri yang berlainan daripada bahasa Indonesia, tetapi tidak mampu menjawab dalam bahasa Jawa sepenuhnya kerana jumlah data latihan yang terhad.
- “Siala jawab dalam bahasa **Jawa**.” – “OK. Input anda bahasa **Jawa**.” – “Saya akan menjawab dalam bahasa **Jawa**.”

Punca utama masalah pemilihan bahasa

- Dakwaan: ChatGPT tidak menganggap bahasa Melayu sebagai bahasa tersendiri malah sebagai dialek bukan standard bagi bahasa Indonesia.
- Fenomena sama:

N

Macam mana nak ke UUM dari KL?

Prom: BM Standard
Lisan/Tidak formal



Untuk pergi ke Universiti Utara Malaysia (UUM) dari Kuala Lumpur, terdapat beberapa pilihan pengangkutan yang boleh anda pertimbangkan. Berikut adalah beberapa cara untuk mencapai destinasi tersebut:

Respons: BM
Standard Formal

Fenomena sama (samb.)

- Jika kita hendak respon dalam bahasa/dialek yang sama dengan prom, kita perlu memberi arahan eksplisit.

N

Sila jawab dalam dialek Kedah.



Oh, nok pegi UUM dari KL, leh amik jalan lebuhraya. Kalau naik kereta, korang boleh amik Lebuhraya Utara-Selatan (E1) terus ke Butterworth-Kulim Expressway (E15). Lepas tu, korang akan lalu tol, ada tol Sungai Dua, pastu sambung jalan sampai ke Sintok, Kedah, yang dekat dengan UUM.

Kalau nak naik bas pulak, boleh gi Terminal Bersepadu Selatan (TBS) kat KL, ada bas Transnasiona, Plusliner, dan Konsortium. Bas-bas ni boleh bawak korang pi Terminal Bas Shahab Perdana kat Sintok,

Kedah. Dari situ, korang boleh amik taksi ke UUM.

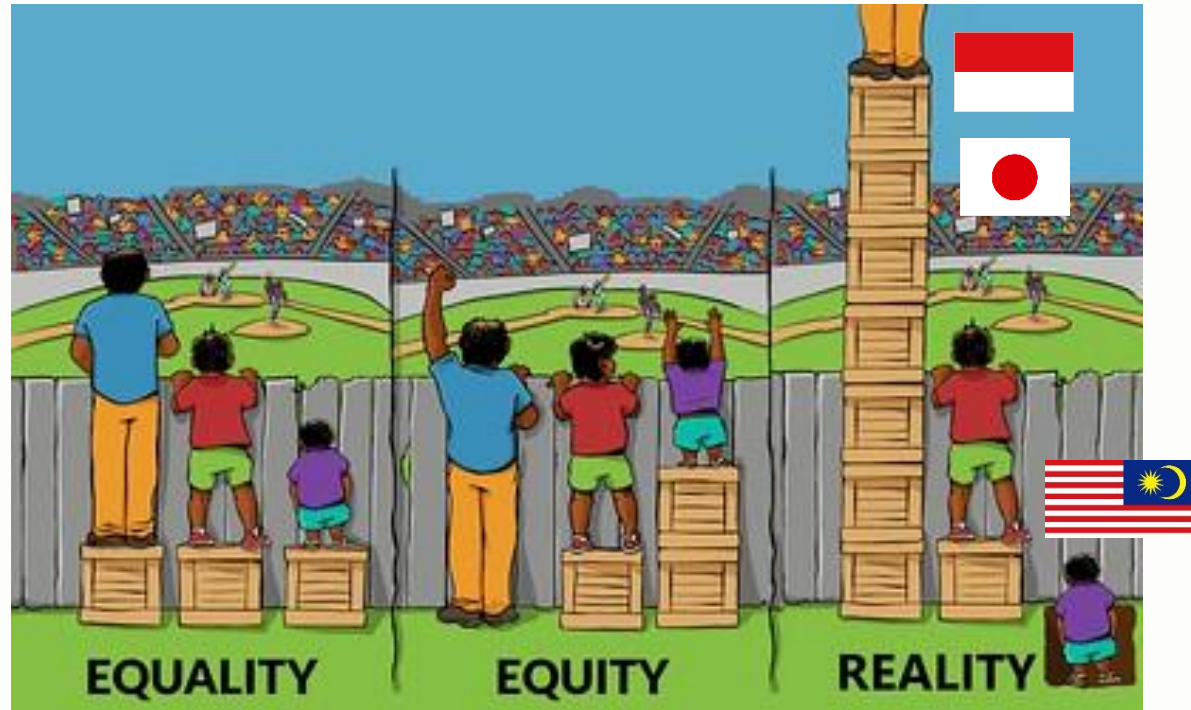
Fenomena sama dengan kes bahasa Jawa (kekurangan data dialek Kedah)

Penilaian masalah pemilihan bahasa

- **Dari segi linguistik:** Boleh dikatakan munasabah kerana bahasa Melayu dan bahasa Indonesia adalah dua kelainan standard bagi bahasa Melayu sebagai bahasa makro (*macrolanguage*).
- **Dari segi sosiopolitik:** Tidak patut berlaku kerana kedua-dua bahasa adalah bahasa tersendiri yang kedudukannya sama rata.

Masalah sosial yang disebabkan oleh ChatGPT (1)

Ketidaksamaan (*inequality*) dan ketaksaksamaan (*inequity*)



<https://interactioninstitute.org/equality-vs-equity-the-gift-that-keeps-on-giving/>

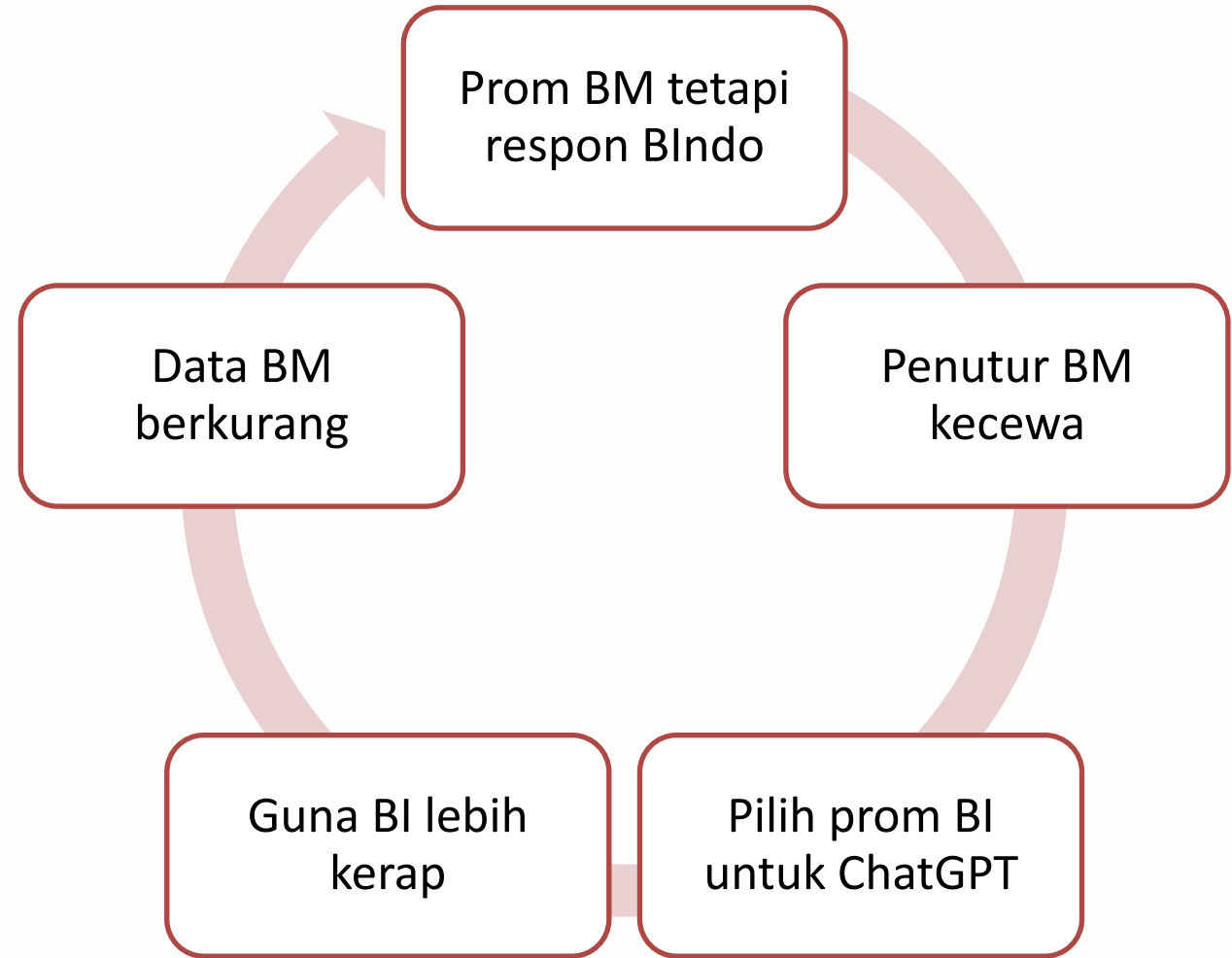
Kedua-dua penutur bahasa Melayu dan bahasa Indonesia boleh mendapat respons dalam bahasa sendiri.

Penutur bahasa Melayu boleh mendapat manfaat yang sama dengan penutur bahasa Indonesia.

Masalah sosial yang disebabkan oleh ChatGPT (2)

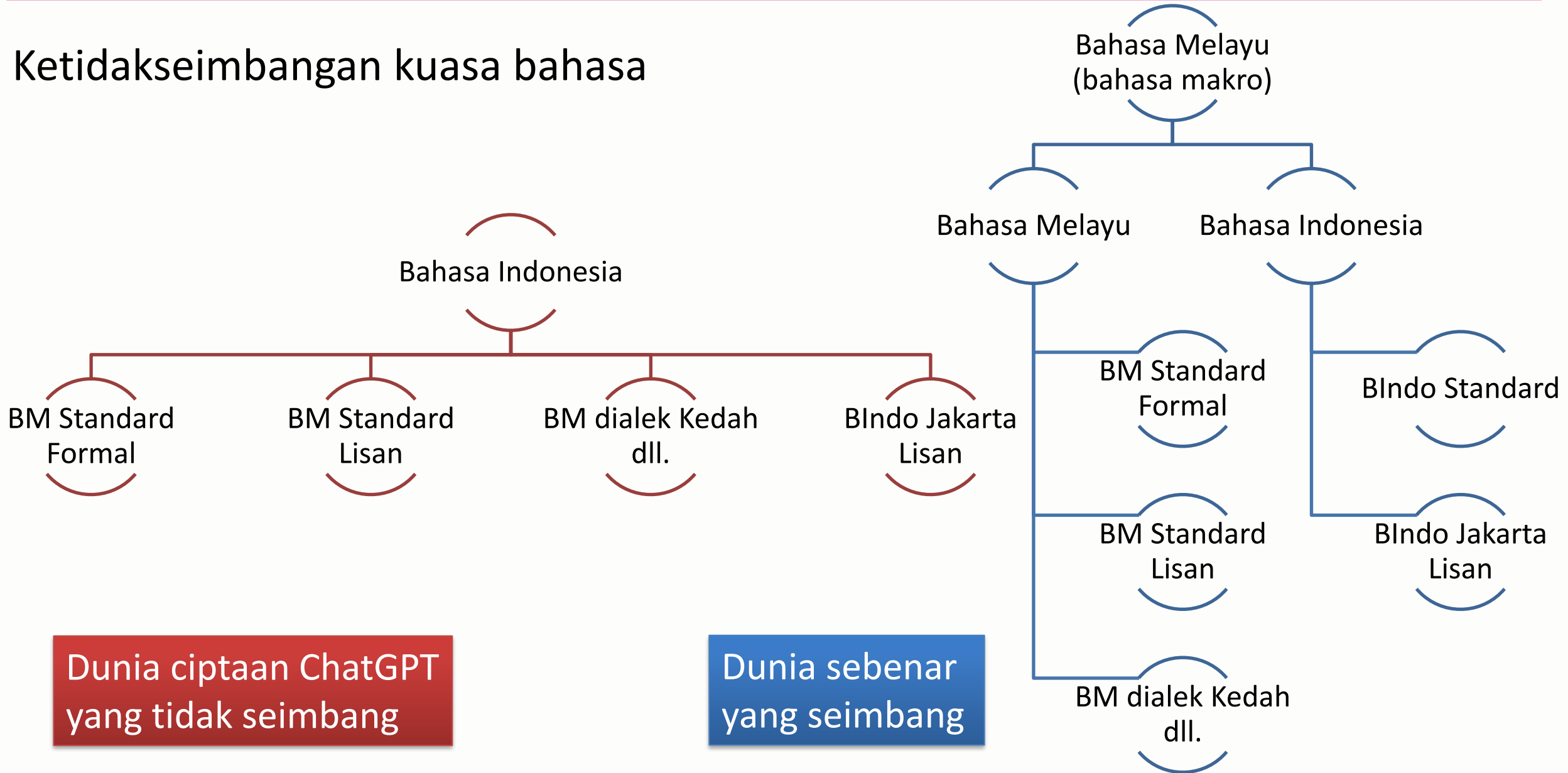
Peralihan bahasa (*language shift*)

- Teknologi maklumat adalah antara ranah yang paling penting yang mempengaruhi vitaliti sesuatu bahasa.
- Bukan sahaja “bahasa dan sastera” tetapi “**bahasa dan teknologi**”



Masalah sosial yang disebabkan oleh ChatGPT (3)

Ketidakseimbangan kuasa bahasa



Cara penyelesaian (1): Tetapan bahasa

- Jika tetapan bahasa diperkenalkan, pengguna tidak perlu memberi prom tambahan seperti “Sila jawab dalam bahasa Melayu, dan bukannya bahasa Indonesia.”
- Cara ini bukan cara ideal kalau hanya penutur bahasa dominan (iaitu bahasa Indonesia) boleh bermanfaat dari fungsi identifikasi bahasa automatik, iaitu tanpa prom tambahan.
- Kebolehan identifikasi bahasa ChatGPT perlu diperbaiki lagi.
cf. Google Translate > manusia > ChatGPT

Cara penyelesaian (2): Senarai bahasa

- Mesti ada senarai bahasa-bahasa yang perlu dibezakan sebagai bahasa tersendiri.
- Senarai tersebut akan mencegah ChatGPT daripada mencipta dunia yang tidak seimbang di mana bahasa Melayu dianggap sebagai dialek bukan standard bagi bahasa Indonesia.
- Sudah diamalkan oleh perkhidmatan penterjemahan automatik seperti Bing Translate dan Google Translate.

Cara penyelesaian (3): Penglibatan kerajaan

- Pemerintah Iceland bekerjasama dengan syarikat OpenAI untuk memperbaiki mutu bahasa Iceland di ChatGPT.

*On the **initiative of the country's President**, HE Guðni Th. Jóhannesson, and with the **help of private industry**, Iceland has partnered with OpenAI to use GPT-4 in the preservation effort of the Icelandic language—and to turn a defensive position into an opportunity to innovate.*

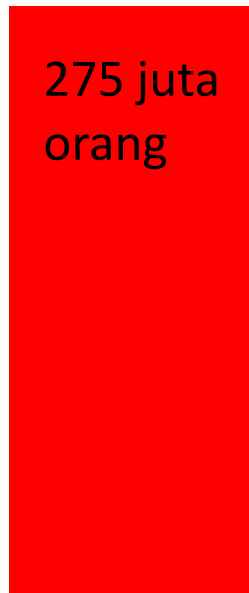
(<https://openai.com/customer-stories/government-of-iceland>)

- Mutu bahasa Melayu sudah cukup baik. Jadi strategi Iceland tidak relevan bagi Malaysia.

Apakah yang boleh dilakukan oleh kerajaan Malaysia?

1. Menggalakkan rakyat dan syarikat untuk menggunakan bahasa Melayu lebih banyak di internet untuk memperbanyakkan data bahasa Melayu.

Keadaan sekarang: Sangat banyak laman web syarikat swasta yang hanya ditulis dalam bahasa Inggeris sahaja.

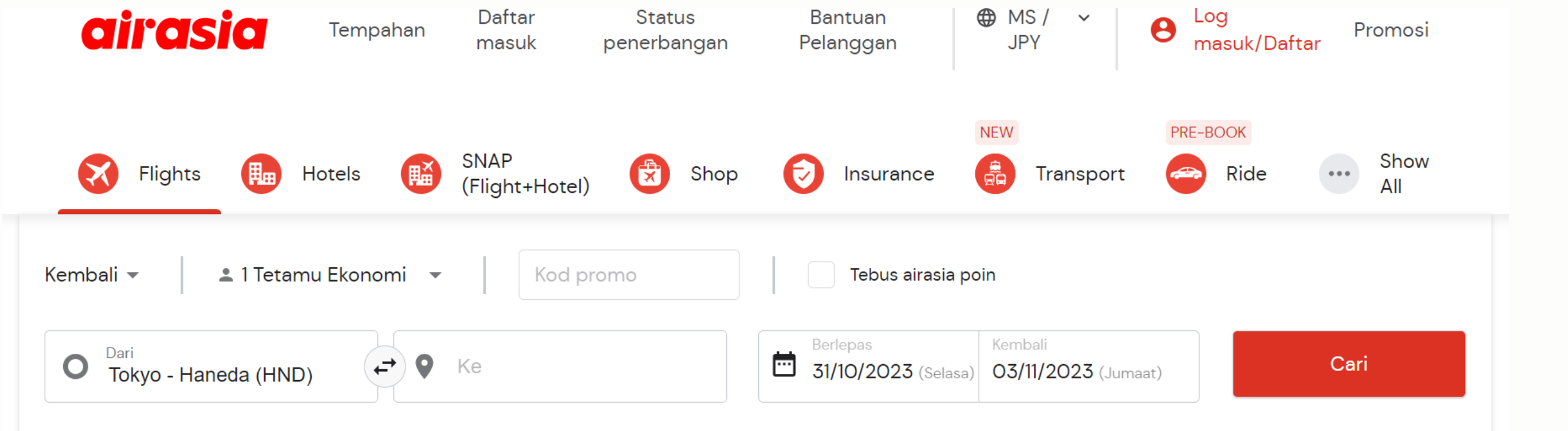
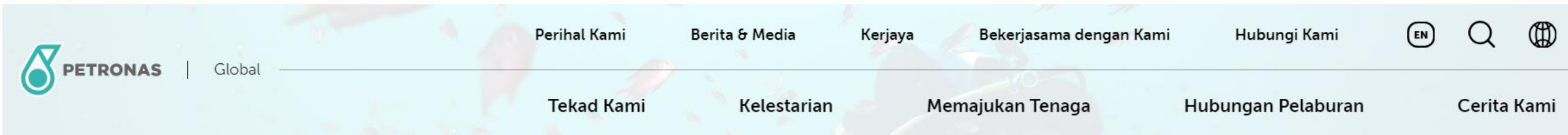


Berapa peratuskah laman BM daripada semua laman web yang dihasilkan oleh rakyat dan syarikat Malaysia?

Berapa peratuskah bidang-bidang yang menggunakan BM sebagai bahasa utama?



Contoh amalan baik/sepatutnya



Apakah yang boleh dilakukan oleh kerajaan Malaysia?

2. Meminta OpenAI dan Common Crawl, iaitu sumber utama dataset Latihan ChatGPT, untuk memperbanyakkan data bahasa Melayu sehingga menjadi sama dengan bahasa Indonesia.

Saiz dataset latihan ChatGPT dari segi perkataan

Bahasa	Perkataan	%
Inggeris	181,014,683,608	92.64708%
Perancis	3,553,061,536	1.81853%
Jepun	217,047,918	0.11109%
Cina (Mandarin)	193,517,396	0.09905%
Indonesia	116,930,321	0.05985%
Melayu	13,389,340	0.00685%
Tamil	5,163,171	0.00264%
Jawa	1,448	0.00000%

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

Laman berdaftar di Common Crawl

500 domain berdaftar yang paling banyak digunakan oleh Common Crawl

(<https://commoncrawl.github.io/cc-crawl-statistics/plots/domains>)

- .jp (24)
u-tokyo.ac.jp, u-kyoto.ac.jp, sakura.ne.jp, exblog.jp, nicovideo.jp, dll.
cf. um.edu.my, usm.my = bahasa Inggeris
- .id (2)
kemdikbud.go.id, republika.co.id
kemdikbud = Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi
- .sg (1)
google.com.sg
- .my (0)

Kesimpulan

- ChatGPT mempunyai masalah pemilihan bahasa. Prom bahasa Melayu kecenderungan dijawab dengan respon bahasa Indonesia.
- Masalah ini boleh dikaitkan dengan kesalahan identifikasi bahasa, tetapi punca utama sebenarnya ialah kerana ChatGPT tidak melayan kedua-dua bahasa secara sama rata.
- Ia berperilaku seolah-olah bahasa Melayu merupakan salah satu dialek bukan standard bagi bahasa Indonesia.
- Masalah pemilihan bahasa ini bukan sahaja masalah teknologi tetapi juga menyebabkan pengaruh sosial yang negatif.
- Masalah tersebut boleh diselesaikan baik secara teknologi mahupun secara sosiopolitik.

Kata-kata penutup

- “Pengembangan AI bertanggungjawab (*responsible AI development*)”
--- untuk syarikat teknologi
- “Pembelian beretika (*ethical consumption*)”
--- untuk pengguna, memilih syarikat yang menyumbang kepada manfaat penutur bahasa Melayu
- “Bahasa dan teknologi” cf. bahasa dan sastera
--- untuk bidang pendidikan, melahirkan sumber manusia yang mahir dalam teknologi bahasa

Rujukan

- Hiroki Nomoto, Shiro Akasegawa dan Asako Shiohara. 2018. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA* 65:47–66.
- Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama dan Chenchen Ding. 2016. Introduction of the Asian Language Treebank. Dlm *Oriental COCOSDA*.