



LLM-Based Tools and Gemini API Integration for Data Scientists

+ Partnerships Hacktiv8

This Program is part of the AI Opportunity Fund: Asia Pacific, in collaboration with AVPN and supported by [Google.org](#) and the Asian Development Bank.



Implementing Generative AI with Gemini

Session 3

Learning Session

Introduction to AI & Generative AI with Gemini

- Pengenalan AI, Generative AI dan LLM
- AI Ethics
- Pengenalan Chatbot
- Hallucinations & Prompting Techniques
- Pengenalan Gemini
- Hands-On

Session 2

Implementing Generative AI with Gemini

- Teknik Prompting Lanjutan
- Konfigurasi Google Gemini
- Pemanggilan Fungsi

Session 3

RAG with LangChain

- Pengenalan RAG
- Vector Database
- Pengenalan LangChain
- Pengenalan Llama
- Hands-On

Session 5

Building Applications with LLM

- Pengenalan Streamlit
- Pembuatan ChatBot LLM dengan Streamlit
- Pengenalan AI Agents
- Mendeploy Aplikasi Streamlit ke Cloud



Training Guidelines

Untuk mengikuti materi pelatihan berikut ada beberapa hal yang harus diperhatikan oleh peserta:

- Peserta diharapkan sudah memiliki pengetahuan dasar tentang Python.
- Peserta diharapkan memperhatikan materi dan mencatat materi atau penjelasan yang disampaikan dari instruktur.
- Instruktur diharapkan dapat menjelaskan materi dengan jelas dan memastikan peserta pelatihan dapat memahami topik yang dibahas.
- Materi yang diajarkan sebagian besar adalah praktik atau menggunakan Google Colab (Jupyter Notebook).

List of Contents

- Advanced Prompting Techniques
- Google Gemini Configurations
- Function Calling



Objectives

- Peserta dapat mengetahui teknik prompting tingkat lanjut.
- Peserta dapat mengetahui dan menerapkan konfigurasi pada Google Gemini.



Advanced Prompting Techniques

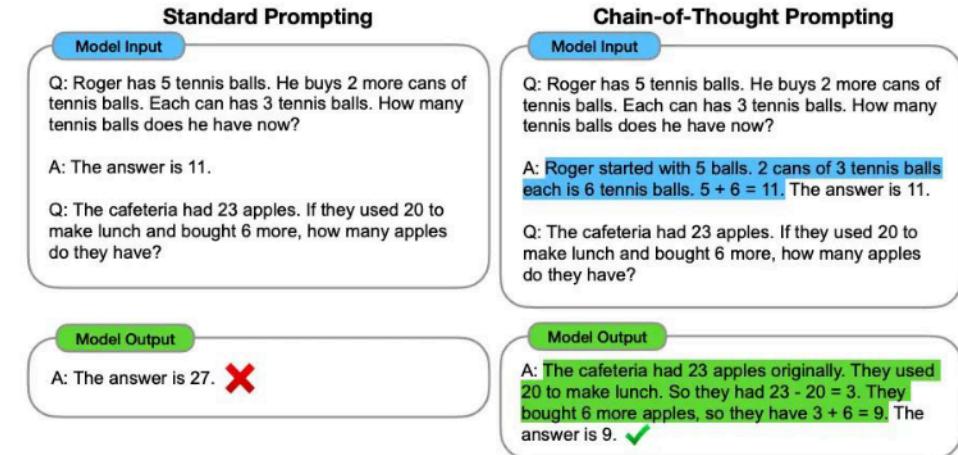
Prompting Techniques

CoT - Chain-of-Thought

Diperkenalkan di Wei et al. (2022), chain-of-thought (CoT) memungkinkan kemampuan penalaran kompleks melalui langkah-langkah penalaran menengah.

Anda dapat menggabungkannya dengan dorongan beberapa langkah untuk mendapatkan hasil yang lebih baik pada tugas yang lebih kompleks yang memerlukan penalaran sebelum merespons.

Untuk menghasilkan respon yang lebih baik, CoT dapat diterapkan dengan memberikan beberapa contoh penalaran (Few-Shot Prompting).



Sumber: <https://www.promptingguide.ai/techniques/cot>

Prompting Techniques

CoT - Chain-of-Thought

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

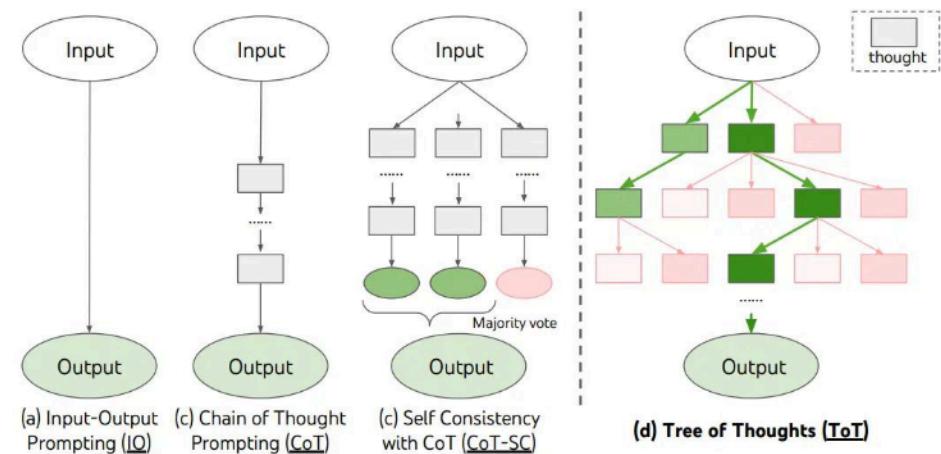
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Prompting Techniques

ToT - Tree of Thoughts

Yao et al. (2023) dan Long (2023) baru-baru ini mengusulkan *Tree of Thoughts* (ToT), sebuah kerangka kerja yang memperluas konsep *chain-of-thought* dan mendorong eksplorasi atas pemikiran-pemikiran yang berfungsi sebagai langkah-langkah perantara dalam pemecahan masalah umum dengan *language model*.

ToT mempertahankan sebuah pohon pemikiran, di mana pemikiran-pemikiran tersebut mewakili urutan bahasa yang koheren yang berfungsi sebagai langkah-langkah perantara untuk menyelesaikan sebuah masalah.



Sumber: <https://www.promptingguide.ai/techniques/tot>

Prompting Techniques

ToT - Tree of Thoughts

Hulbert (2023) telah mengusulkan Tree-of-Thought Prompting, yang menerapkan konsep utama dari kerangka kerja ToT sebagai teknik prompt sederhana, yang membuat LLM mengevaluasi pemikiran antara dalam satu prompt tunggal. Contoh prompt ToT adalah:

Imagine three different experts are answering this question.

*All experts will write down 1 step of their thinking,
then share it with the group.*

Then all experts will go on to the next step, etc.

If any expert realises they're wrong at any point then they leave.

The question is...

Prompting Techniques

ReAct

Yao et al., 2022 memperkenalkan sebuah kerangka kerja bernama **ReAct**, di mana **LLM** digunakan untuk menghasilkan jejak penalaran atau berpikir (reasoning traces) dan tindakan spesifik tugas (task-specific actions) secara bergantian.

ReAct terinspirasi oleh sinergi antara "**acting**" dan "**reasoning**" yang memungkinkan manusia untuk mempelajari tugas baru dan membuat keputusan atau penalaran. Teknik ini dapat mengatasi kekurangan teknik **CoT** yang memiliki **ketidakmampuannya untuk mengakses dunia luar atau tidak dapat memperbarui pengetahuannya**, yang dapat menyebabkan masalah seperti **halusinasi**.

Reasoning traces memungkinkan model untuk membuat, melacak, dan memperbarui rencana tindakan, serta menangani pengecualian atau masalah yang muncul. **Action step** memungkinkan model untuk berinteraksi dengan dan mengambil informasi dari sumber eksternal, seperti basis pengetahuan atau lingkungan.

Prompting Techniques

ReAct

ReAct meminta LLMs untuk menghasilkan **jejak penalaran verbal** dan **tindakan** untuk suatu tugas. Hal ini memungkinkan sistem untuk melakukan penalaran dinamis untuk membuat, memelihara, dan menyesuaikan rencana tindakan, sekaligus memungkinkan **interaksi dengan lingkungan eksternal** (misalnya, Wikipedia) untuk menyertakan informasi tambahan dalam proses penalaran.

Untuk lebih detail memahami ReAct, Anda dapat melihat [di sini](#)

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the `Front Row media center` program ...

Thought 2: Apple Remote was originally designed to control the `Front Row media center` program. I need to search `Front Row` next and find what other device can control it.

Act 2: `Search[Front Row]`

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search `Front Row (software)`.

Act 3: `Search[Front Row (software)]`

Obs 3: Front Row is a discontinued media center software ...

Thought 4: `Front Row (software)` is controlled by an `Apple Remote` or the `keyboard function keys`. So the answer is `keyboard function keys`.

Act 4: `Finish[keyboard function keys]`



Prompting Techniques

Macam-Macam Teknik Prompting

Ada banyak sekali teknik prompting yang dapat anda pelajari selain dari yang telah dijelaskan. Berikut Adalah beberapa teknik lainnya:

- Self-Consistency
- Automatic Reasoning and Tool-use (ART)
- Reflexion
- Directional Stimulus Prompting

Selain teknik tersebut terdapat banyak lagi teknik yang bisa kita gunakan. Anda dapat melihat berbagai macam teknik tersebut pada link berikut:

<https://www.promptingguide.ai/techniques>

Maju Bareng AI

Presence - AI for Data Scientist

Buat yang belum absen,
absen yuk! WAJIB



<https://bit.ly/absensi-data-mba>



Konfigurasi Gemini

Konfigurasi Gemini

Macam-Macam Teknik Prompting

Setiap perintah yang Anda kirim ke model menyertakan parameter yang mengontrol cara model menghasilkan respons. Anda dapat menggunakan GenerationConfig parameter ini untuk mengonfigurasinya.

Macam-macam Parameter:

- max_output_tokens
- temperature
- top_k
- top_p

Konfigurasi Gemini

Temperature dan Max Output

Tokens

- **Temperature** merupakan parameter yang mengontrol tingkat keacakan dalam pemilihan token dan kreativitas teks yang dihasilkan.
- Skor Temperature berada pada rentang 0-2.
- Temperature yang **lebih tinggi** menyebabkan:
 - Hasil yang **lebih beragam** dan kreatif
 - Terkadang teks yang didapat mungkin kurang relevan.
- Temperature yang **lebih rendah** menyebabkan:
 - Teks yang **kurang kreatif** dan beragam.
 - Temperature 0 bersifat deterministik, artinya respons dengan probabilitas tertinggi selalu dipilih.
- **Max Output Tokens** Menentukan jumlah token maksimum yang dapat dihasilkan dalam respons. Untuk model Gemini, **satu token** terdiri dari sekitar **empat karakter**. 100 token setara dengan sekitar 60-80 kata bahasa Inggris.

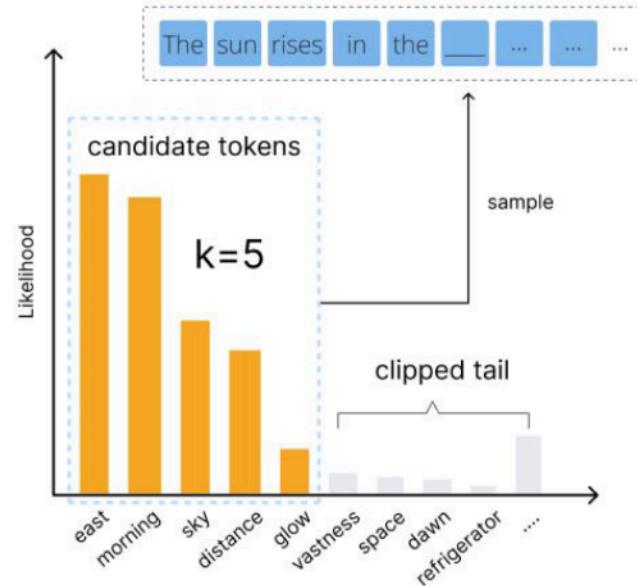
Konfigurasi Gemini

Top-K

Top K mengatur dalam pengambilan sampel token dengan probabilitas tertinggi hingga jumlah token yang ditentukan tercapai.

Jika top K=1 , ini berarti hanya mengambil 1 sampel token dengan probabilitas paling tinggi. Ini bisa mengakibatkan response yang kurang kreatif (Greedy decoding).

Nilai top K umumnya pada rentang 1-100.



Sumber: <https://www.promptingguide.ai/techniques/tot>

Konfigurasi Gemini

Top-P

Top P mengatur dalam pengambilan sampel token dengan skor probabilitas tertinggi hingga jumlah skor mencapai nilai ambang batas yang ditentukan. Top P umumnya memiliki nilai dengan rentang 0 hingga 1.

Top-k sampling (set k to 2)

The dog is

[Top-k]	[Token]	[Top-p (p: Probability)]
Top-1	very	0.81
Top-2	good	0.14
Top-3	bad	0.02
Top-4	wrong	0.002
...
Top-4899	was	0.0001
Top-4900	will	0.0001
...

(Sampling list)

Top-p sampling (set p to 0.96)

The dog is

[Top-k]	[Token]	[Top-p (p: Probability)]
Top-1	very	0.81
Top-2	good	0.14
Top-3	bad	0.02
Top-4	wrong	0.002
...
Top-4899	was	0.0001
Top-4900	will	0.0001
...

Cumulative probability exceeds 0.96
($0.81+0.14+0.02 > 0.96$)

(Sampling list)

Sumber: https://www.researchgate.net/figure/Process-of-top-k-and-top-p-sampling_fig6_376832794



Function Calling

Function Calling

Apa itu Function Calling?

Function calling adalah teknik yang memungkinkan model bahasa besar (LLM) seperti Gemini (Google) atau Llama (Meta) berinteraksi dengan sistem eksternal (API, database, dll.) dengan menghasilkan output terstruktur.

Contoh function calling dapat dilihat seperti pada diagram di samping, dimana LLM dapat digunakan untuk melakukan pencarian pada Google Search atau memesan tiket pesawat. Semua dapat dilakukan dengan beberapa kalimat perintah saja.

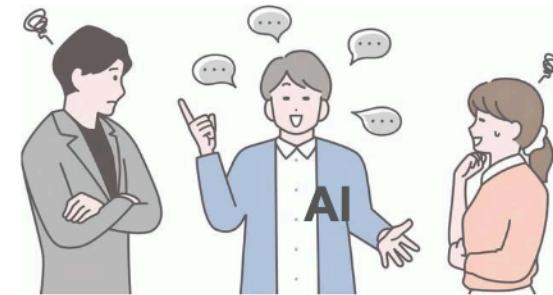


Function Calling

Manfaat Function Calling

Sebelum ada function calling, interaksi manusia dengan LLM hanya sebatas percakapan biasa, LLM memberikan informasi yang dibutuhkan namun tetap manusia lah yang melakukan "aksi"nya. Setelah adanya function calling interaksi dapat berupa aksi nyata yang langsung dilakukan oleh LLM.

Dengan function calling, LLM tidak hanya dapat menghasilkan teks namun juga dapat melakukan action sesuai dengan apa yang dibutuhkan berdasarkan input yang diberikan oleh pengguna.



Sebelum ada function calling
Setelah ada function calling



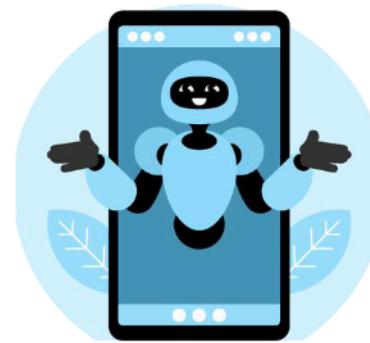
Function Calling

Contoh Implementasi



Virtual Assistant

Virtual assistant pada ponsel dapat melakukan berbagai tugas seperti membuka aplikasi atau memutar lagi hanya dengan beberapa kalimat perintah.



Chatbot

Saat ini beberapa perusahaan besar di Indonesia sudah memanfaatkan chatbot untuk berinteraksi dengan customernya. Biasanya, pada chatbot ini juga dilengkapi dengan function calling seperti pengecekan data customer atau melakukan transaksi.

Maju Bareng AI

Presence - AI for Data Scientist

Buat yang belum absen,
absen yuk! WAJIB



<https://bit.ly/absensi-data-mba>



Latihan Gemini

Hands-on

Latihan

Berikut adalah link google colab yang dapat anda gunakan sebagai referensi untuk latihan untuk konfigurasi Google Gemini menggunakan Python:

- <https://colab.research.google.com/drive/1b-Lq2dy2SFbP3Zqy56JD9RHNTkcFIkdm?usp=sharing>



Quiz

LLM-Based Tools and Gemini API Integration for Data Scientists

Quiz Session 3



<https://bit.ly/quizdata3>

Due Date:
Hari ini, 23.59 WIB

Ingat, absensi juga dihitung
dari pengisian Quiz.

Certificate from Hacktiv8

Phase	Task	Due Date
Session 1	Presence	Session 1
Session 2	Presence & Quiz 1	Session 2
Session 3	Presence & Quiz 2	Session 3
Session 4	Presence & Quiz 3	Session 4
Session 5	Presence & Quiz 4	Session 5
Final Project	Form Submission: Final Project	H+2 Session 5

✨ Special Reward! ✨

Di akhir program, akan ada merchandise yang dibagikan untuk peserta yang beruntung yang telah menyelesaikan seluruh rangkaian program* 🎁

*Satu pemenang untuk satu training topics.



Thank You

Hacktiv8

www.hacktiv8.com

(021) 8067 5787
halo@hacktiv8.com