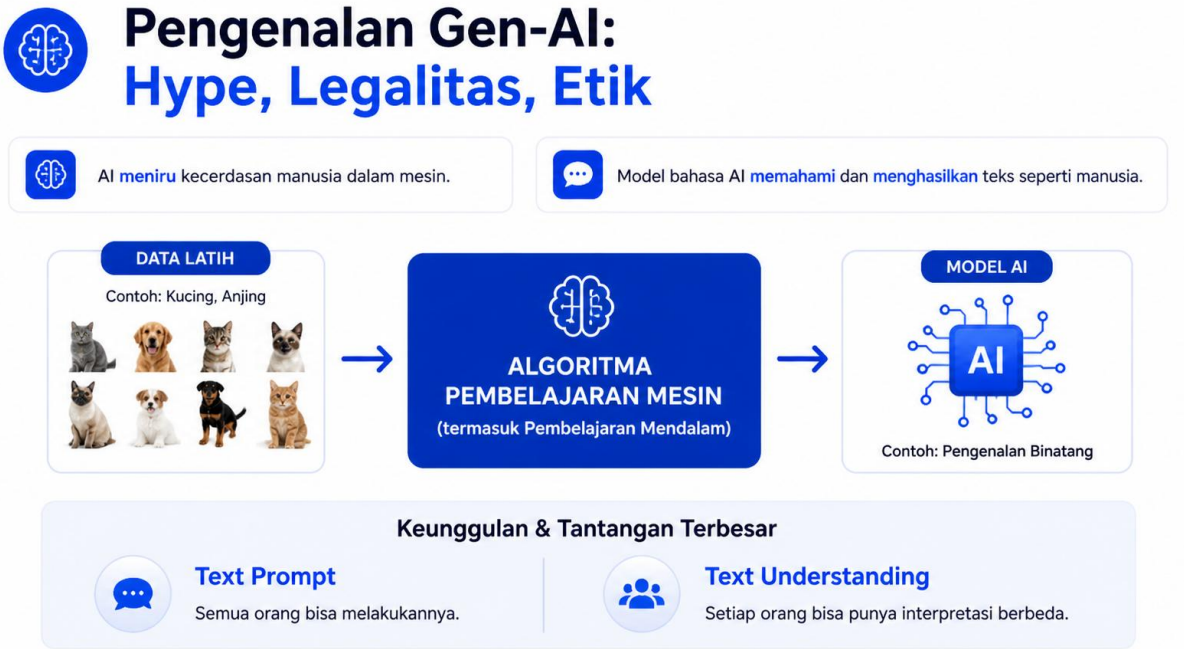
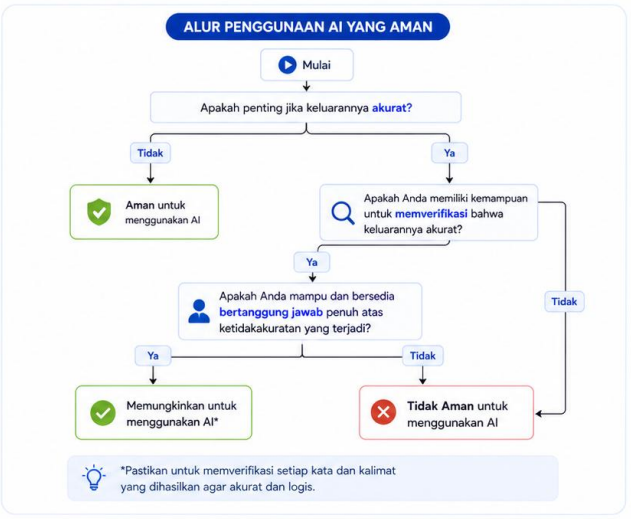

## Materi

- Pengantar Gen-AI dan Workspace Chat-GPT
- Pengantar SLR
- Demo Prompting SLR
- Prompting SLR dengan CustomGPT + Rayyan



# Aspek Legal & Etis: Pengguna AI Bertanggungjawab

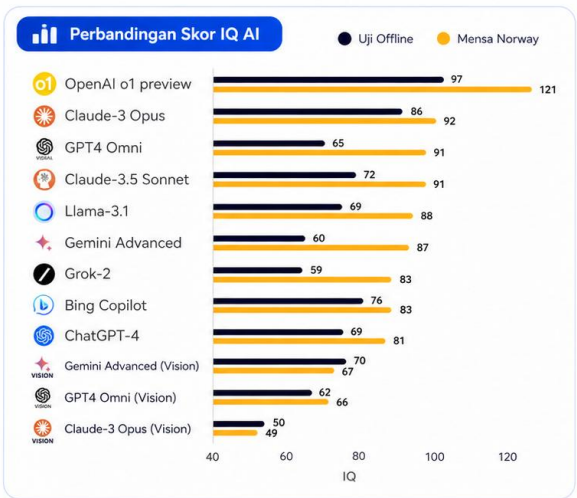




Panduan UNESCO terkait
Alur Penggunaan ChatGPT yang Aman

- **Verifikasi** — Selalu cek kebenaran informasi.
- **Tanggung Jawab** — Anda bertanggung jawab atas hasil yang digunakan.
- **Etika** — Gunakan untuk kebaikan dan sesuai aturan.
- **Privasi** — Jaga data pribadi dan informasi sensitif.

# Inteligensi AI

Data: Oktober 2024





Model **OpenAI o1 preview** meraih skor **120** dalam uji IQ Mensa Norway.

RESEARCH ARTICLES

MACHINE LEARNING

## Performance of a large language model on the reasoning tasks of a physician

Peter G. Brodeur[1†], Thomas A. Buckley[2†], Zahir Kanjee[3], Ethan Goh[4,5], Evelyn Bin Ling[6], Priyank Jain[7], Stephanie Cabral[8,7], Raja-Elie Abdulnour[9], Adrian D. Haimovich[9], Jason A. Freed[10], Andrew Olson[11], Daniel J. Morgan[12,13], Jason Hom[14], Robert Gallo[4,15], Liam G. McCoy[16,17], Hoadi Mombini[18], Christopher Lucas[1], Misha Fotoohi[1], Matthew Gwiazdon[1], Daniele Restifo[1], Daniel Restrepo[19], Eric Horvitz[20,21], Jonathan Chen[14,22]‡, Arjun K. Manrai[23]*‡, Adam Rodman[24]*‡

More than 65 years ago, complex clinical diagnostic reasoning cases were introduced as the gold standard for the evaluation of expert medical computing systems, a standard that has held ever since. In this study, we report the results of a physician evaluation of a large language model (LLM) on challenging clinical cases across five experiments with a baseline of hundreds of physicians. We then report a real-world study comparing human expert and artificial intelligence (AI) second opinions in randomly selected patients in the emergency room of a major tertiary academic medical center. In all experiments, the LLM outperformed physician baselines and displayed continued improvement from prior generations of AI clinical decision support. Our study suggests that LLMs have eclipsed most benchmarks of clinical reasoning, motivating the urgent need for prospective trials.

Artificial intelligence (AI) diagnostic support tools have been studied since the 1950s, after a landmark paper was published in *Science* by Ledley and Lusted (*1*), who advocated for case-based benchmarks as an evaluation standard, a standard that has held for more than the past half century (*2*–*6*). In particular, the *New England Journal of Medicine* (*NEJM*) clinicopathological case conference series has been seen as an aspirational goal post, tested by every differential diagnosis generator spanning primitive Bayesian systems, symbolic rules–based systems, and natural-language symptom checkers. Recently, large language models (LLMs) have consistently outperformed older models on these challenging cases, mirroring their performance in professional licensing exams, mathematics questions, software engineering, and engineering problems (*7*–*12*).

However, recent studies of LLMs in medicine have focused on narrow diagnostic tasks or on curated and constrained clinical vignettes (*7*, *13*, *14*). More importantly, most studies of LLMs for diagnosis and management to date have lacked human physician baselines. This was justifiable in previous generations of technology because of the overall poor performance of prior computational models on benchmarks. Given rapid improvement in LLMs and increasing "benchmark saturation," it is necessary to establish human baselines and study clinically grounded tasks. Here, we comprehensively evaluated the diagnostic and management reasoning capabilities of an advanced LLM (OpenAI o1 series) across several diagnostic and management reasoning tasks with baseline performance from hundreds of physicians. We further studied LLM second opinions in a blinded fashion against an expert physician baseline on randomly selected patients in a major academic tertiary care emergency department in Boston, Massachusetts.

### Results

#### Quality of differential diagnoses and testing plans on *NEJM* clinicopathological conferences

We first evaluated o1-preview using the clinicopathologic conferences (CPCs) published by the *NEJM*, a standard for the evaluation of differential generators since the 1959s (*1*). There was substantial agreement between the two physicians evaluating the quality of o1-preview's differential diagnosis [agreement on 120/163 cases (84%), inter-rater reliability ($\kappa$) = 0.68]. o1-preview included the correct diagnosis in its differential in 78.3% of cases [95% confidence interval (CI), 70.7 to 84.8%] (Fig. 1). The first diagnosis suggested was the correct diagnosis in 52% of cases (95% CI, 44 to 61%). When expanding to also include potentially helpful or very close diagnoses, o1-preview was accurate on 97.9% (95% CI, 94.0 to 99.8%) of cases (Fig. 1A). We did not find evidence of a significant difference in performance before and after the premaining cutoff date for o1-preview (79.8% accuracy before, 73.5% accuracy after, $P$ = 0.59; table S1). In a subset of 101 cases from a prior study (*8*), o1-preview outperformed a human physician baseline in both top-1 and top-10 accuracy (table S1). On 70 cases used to evaluate GPT-4 in a prior study (*7*), o1-preview produced a response with the exact or a very close diagnosis in 88.6% of cases, compared with 72.9% of cases by GPT-4 ($P$ = 0.03; Fig. 2B). Overall, o1-preview and GPT-4 performed identically on 66/70 (69.6%) of cases, o1-preview outperformed GPT-4 on 17/70 (24.3%) of cases, and GPT-4 outperformed o1-preview on 5/70 (7.1%) of cases (Fig. S1).

We next evaluated the ability of o1-preview to select the next diagnostic test in the *NEJM* CPCs for a subset of 136 cases. Two physicians scored the suggested test plan produced by o1-preview [agreement on 113/130 cases (87%), $\kappa$ = 0.28] with respect to the actual management of the patient described in the CPC. The proportion of agreements was high, but the $\kappa$ was low as a result of severe class imbalance. In 87.5% of cases, o1-preview selected the correct test to order; in another 11% of cases, the chosen testing plan was judged by the two physicians to be helpful; and in 1.5% of cases, it would have been unhelpful (Fig. 2C). Inference costs and examples of model outputs are shown in tables S3 to S5.

#### Presentation of reasoning in *NEJM* Healer diagnostic cases

We used 20 clinical reasoning cases from the *NEJM* Healer curriculum (*15*) that were also evaluated in a prior study using GPT-4 (*16*). *NEJM* Healer cases are virtual patient encounters designed for the assessment of clinical reasoning (*15*). There was substantial agreement of Revised-IDEA (R-IDEA) scores—a validated 10-point scale for evaluating four core domains of documenting clinical reasoning (*17*)—between the two physicians [agreement on 79/80 (99%) cases, $\kappa$ = 0.89]. For 78/80 of the cases, o1-preview achieved a perfect R-IDEA score, significantly

[1]Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Stanford Division of Computational Medicine, Stanford University, Stanford, CA, USA. [4]Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [5]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Internal Medicine, Cambridge Health Alliance, Cambridge, MA, USA. [7]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA. [8]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [10]Department of Hematology-Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [11]Department of Hospital Medicine, University of Minnesota Medical School, Minneapolis, MN, USA. [12]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA. [13]Veterans Affairs Maryland Healthcare System, Baltimore, MD, USA. [14]Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA, USA. [15]Division of Hospital Medicine at Zuckerberg San Francisco General Hospital, University of California, San Francisco, San Francisco, CA, USA. [16]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [17]Division of Neurology, University of Alberta, Edmonton, AB, Canada. [18]Technology and Innovation Group, Beth Israel Lahey Health, Boston, MA, USA. [19]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [20]Microsoft, Redmond, WA, USA. [21]Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA. [22]Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, CA, USA. *Corresponding author. Email: arjun_manrai@hms.harvard.edu (A.K.M.); arodman@bidmc.harvard.edu (A.R.) †These authors contributed equally to this work. ‡These authors contributed equally to this work.



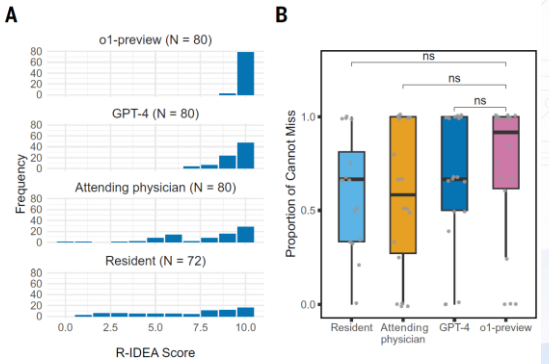



**Fig. 3. Comparison of o1-preview, GPT-4, and physicians for clinical diagnostic reasoning.** (**A**) Distribution of 312 R-IDEA scores stratified by respondents on 20 *NEJM* Healer cases. (**B**) Box plot of the proportion of cannot-miss diagnoses included in differential diagnosis for the initial triage presentation. The total sample size in this figure is 70, with 18 responses from attending physicians, GPT-4, and o1-preview, and 16 responses from residents. Two cases were excluded because the cannot-miss diagnoses could not be identified. ns, not significant.

RESEARCH ARTICLES

MACHINE LEARNING

## Performance of a large language model on the reasoning tasks of a physician

Peter G. Brodeur[1†], Thomas A. Buckley[2†], Zahir Kanjee[3], Ethan Goh[4,5], Evelyn Bin Ling[6], Priyank Jain[7], Stephanie Cabral[8,7], Raja-Elie Abdulnour[9], Adrian D. Haimovich[9], Jason A. Freed[10], Andrew Olson[11], Daniel J. Morgan[12,13], Jason Hom[14], Robert Gallo[4,15], Liam G. McCoy[16,17], Hoadi Mombini[18], Christopher Lucas[1], Misha Fotoohi[1], Matthew Gwiazdon[1], Daniele Restifo[1], Daniel Restrepo[19], Eric Horvitz[20,21], Jonathan Chen[14,22]‡, Arjun K. Manrai[23]*‡, Adam Rodman[24]*‡

More than 65 years ago, complex clinical diagnostic reasoning cases were introduced as the gold standard for the evaluation of expert medical computing systems, a standard that has held ever since. In this study, we report the results of a physician evaluation of a large language model (LLM) on challenging clinical cases across five experiments with a baseline of hundreds of physicians. We then report a real-world study comparing human expert and artificial intelligence (AI) second opinions in randomly selected patients in the emergency room of a major tertiary academic medical center. In all experiments, the LLM outperformed physician baselines and displayed continued improvement from prior generations of AI clinical decision support. Our study suggests that LLMs have eclipsed most benchmarks of clinical reasoning, motivating the urgent need for prospective trials.

Artificial intelligence (AI) diagnostic support tools have been studied since the 1950s, after a landmark paper was published in *Science* by Ledley and Lusted (*1*), who advocated for case-based benchmarks as an evaluation standard, a standard that has held for more than the past half century (*2*–*6*). In particular, the *New England Journal of Medicine* (*NEJM*) clinicopathological case conference series has been seen as an aspirational goal post, tested by every differential diagnosis generator spanning primitive Bayesian systems, symbolic rules–based systems, and natural-language symptom checkers. Recently, large language models (LLMs) have consistently outperformed older models on these challenging cases, mirroring their performance in professional licensing exams, mathematics questions, software engineering, and engineering problems (*9*–*12*).

However, recent studies of LLMs in medicine have focused on narrow diagnostic tasks or on curated and constrained clinical vignettes (*7*, *13*, *14*). More importantly, most studies of LLMs for diagnosis and management to date have lacked human physician baselines. This was justifiable in previous generations of technology because of the overall poor performance of prior computational models on benchmarks. Given rapid improvement in LLMs and increasing "benchmark saturation," it is necessary to establish human baselines and study clinically grounded tasks. Here, we comprehensively evaluated the diagnostic and management reasoning capabilities of an advanced LLM (OpenAI o1 series) across several diagnostic and management reasoning tasks with baseline performance from hundreds of physicians. We further studied LLM second opinions in a blinded fashion against an expert physician baseline on randomly selected patients in a major academic tertiary care emergency department in Boston, Massachusetts.

### Results

#### Quality of differential diagnoses and testing plans on *NEJM* clinicopathological conferences

We first evaluated o1-preview using the clinicopathologic conferences (CPCs) published by the *NEJM*, a standard for the evaluation of differential generators since the 1959s (*1*). There was substantial agreement between the two physicians evaluating the quality of o1-preview's differential diagnosis [agreement on 120/163 cases (84%), inter-rater reliability ($\kappa$) = 0.68]. o1-preview included the correct diagnosis in its differential in 78.3% of cases [95% confidence interval (CI), 70.7 to 84.8%] (Fig. 1). The first diagnosis suggested was the correct diagnosis in 52% of cases (95% CI, 44 to 61%). When expanding to also include potentially helpful or very close diagnoses, o1-preview was accurate on 97.9% (95% CI, 94.0 to 99.8%) of cases (Fig. 1A). We did not find evidence of a significant difference in performance before and after the premaining cutoff date for o1-preview (79.8% accuracy before, 73.5% accuracy after, $P$ = 0.59; table S1). In a subset of 101 cases from a prior study (*8*), o1-preview outperformed a human physician baseline in both top-1 and top-10 accuracy (table S1). On 70 cases used to evaluate GPT-4 in a prior study (*7*), o1-preview produced a response with the exact or a very close diagnosis in 88.6% of cases, compared with 72.9% of cases by GPT-4 ($P$= 0.03; Fig. 2B). Overall, o1-preview and GPT-4 performed identically on 66/70 (69.6%) of cases, o1-preview outperformed GPT-4 on 17/70 (24.3%) of cases, and GPT-4 outperformed o1-preview on 5/70 (7.1%) of cases (Fig. S1).

We next evaluated the ability of o1-preview to select the next diagnostic test in the *NEJM* CPCs for a subset of 136 cases. Two physicians scored the suggested test plan produced by o1-preview [agreement on 113/130 cases (87%), $\kappa$ = 0.28] with respect to the actual management of the patient described in the CPC. The proportion of agreements was high, but the $\kappa$ was low as a result of severe class imbalance. In 87.5% of cases, o1-preview selected the correct test to order; in another 11% of cases, the chosen testing plan was judged by the two physicians to be helpful; and in 1.5% of cases, it would have been unhelpful (Fig. 2C). Inference costs and examples of model outputs are shown in tables S3 to S5.

#### Presentation of reasoning in *NEJM* Healer diagnostic cases

We used 20 clinical reasoning cases from the *NEJM* Healer curriculum (*15*) that were also evaluated in a prior study using GPT-4 (*16*). *NEJM* Healer cases are virtual patient encounters designed for the assessment of clinical reasoning (*15*). There was substantial agreement of Revised-IDEA (R-IDEA) scores—a validated 10-point scale for evaluating four core domains of documenting clinical reasoning (*17*)—between the two physicians [agreement on 79/80 (99%) cases, $\kappa$ = 0.89]. For 78/80 of the cases, o1-preview achieved a perfect R-IDEA score, significantly

[1]Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Stanford Division of Computational Medicine, Stanford University, Stanford, CA, USA. [4]Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [5]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Internal Medicine, Cambridge Health Alliance, Cambridge, MA, USA. [7]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA. [8]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [10]Department of Hematology-Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [11]Department of Hospital Medicine, University of Minnesota Medical School, Minneapolis, MN, USA. [12]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA. [13]Veterans Affairs Maryland Healthcare System, Baltimore, MD, USA. [14]Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA, USA. [15]Division of Hospital Medicine at Zuckerberg San Francisco General Hospital, University of California, San Francisco, San Francisco, CA, USA. [16]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [17]Division of Neurology, University of Alberta, Edmonton, AB, Canada. [18]Technology and Innovation Group, Beth Israel Lahey Health, Boston, MA, USA. [19]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [20]Microsoft, Redmond, WA, USA. [21]Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA. [22]Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, CA, USA. *Corresponding author. Email: arjun_manrai@hms.harvard.edu (A.K.M.); arodman@bidmc.harvard.edu (A.R.) †These authors contributed equally to this work. ‡These authors contributed equally to this work.



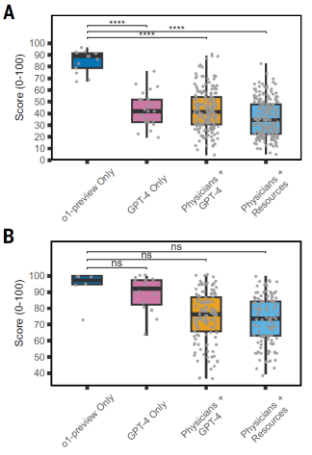



**Fig. 4. Comparison of o1-preview, GPT-4, and physicians for management and diagnostic reasoning.** (**A**) Box plot of normalized management reasoning points by LLMs and physicians on *Grey Matters* Management Cases. Five cases were included. We generated three o1-preview responses for each case. The prior study collected five GPT-4 responses to each case, 178 completed cases from 46 physicians with access to GPT-4, and 197 completed cases from 46 physicians with access to conventional resources. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$. (**B**) Box plot of normalized diagnostic reasoning points by LLMs and physicians. Six diagnostic challenges were included. We generated one o1-preview response for each case. The prior study collected three GPT-4 responses to all cases, 125 cases completed by 25 physicians with access to GPT-4, and 119 cases completed by 25 physicians with access to conventional resources.

# Pengenalan Chat GPT

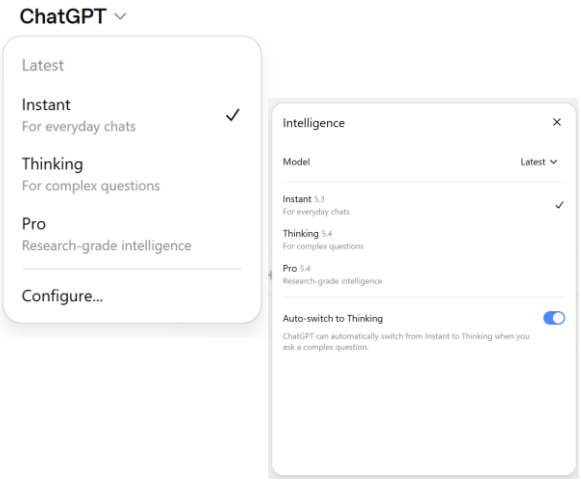
## Jenis Akun dan Keamanan

- Data Tidak Aman: (Free, Plus)
- Data Aman: Business, Enterprise, Pro

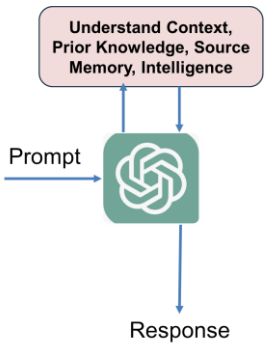
## GPT Ecosystem

- Canvas: Kolaborasi dengan AI
- Project: management
- Membagi Chat (Sesama member GPT Teams)
- Custom-GPT

## Model/Algoritma:




## Cara kerja: Internal vs external knowledge (*Context*)




Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

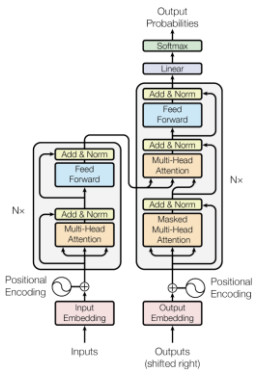



Figure 1: The Transformer - model architecture.

## Pilih Model GPT yang Tepat

**Gunakan model sesuai kompleksitas tugas, kedalaman analisis, dan kebutuhan waktu Anda.**

| Aspek | STANDARD | ADVANCED | DEEP RESEARCH |
|---|---|---|---|
| **Model** | Instant, Thinking-Standard | Thinking-Extended, Pro | Deep Research |
| **Deskripsi** | Tugas sehari-hari dan kebutuhan cepat | Analisis mendalam dan tugas kompleks | Riset komprehensif dan keputusan berbasis data |
| **Kecepatan** | Cepat | Sedang | Lebih lama |
| **Kedalaman Analisis** | Standar | Mendalam | Sangat mendalam |
| **Akurasi** | Baik | Tinggi | Sangat tinggi |
| **Tugas Ideal** | Aktivitas sehari-hari | Tugas kompleks & strategis | Riset & insight kritis |
| **Penggunaan** | Efisien | Profesional | Analisis tingkat lanjut |

## Keterbatasan Model Sederhana LLM (Model: Instant, Thinking-Standard)

- Tidak mendalam/spesifik
- Tanpa sumber/referensi
- Kesulitan pada konsep kompleks
- Sering keliru memilik fakta/fiksi
- Berhalusinasi

*"Fully guided AI as writing assistance"*

**Strategic Prompting**

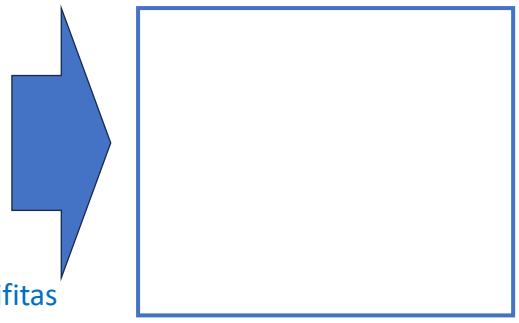
**Prompt Engineering**

**Prompting Framework**

**Memberi ruang bagi pengguna untuk berkreasi**

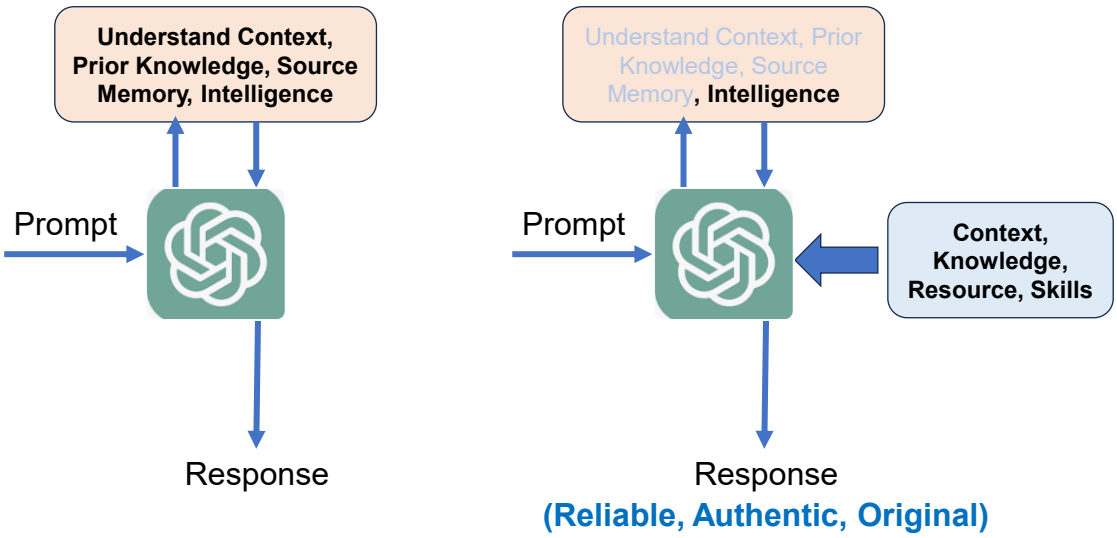
## Keterbatasan Model Lanjut LLM (Pro, Thinking Extended)

- ~~Tidak mendalam/spesifik~~
- ~~Tanpa sumber/referensi~~
- ~~Kesulitan pada konsep kompleks~~
- ~~Sering keliru memilik fakta/fiksi~~
- ~~Berhalusinasi~~
- Sedikit ruang untuk berkreasi/kreatifitas
- Kualitas Respond >>> Kapasitas Prompter (**AI lebih dominan**)

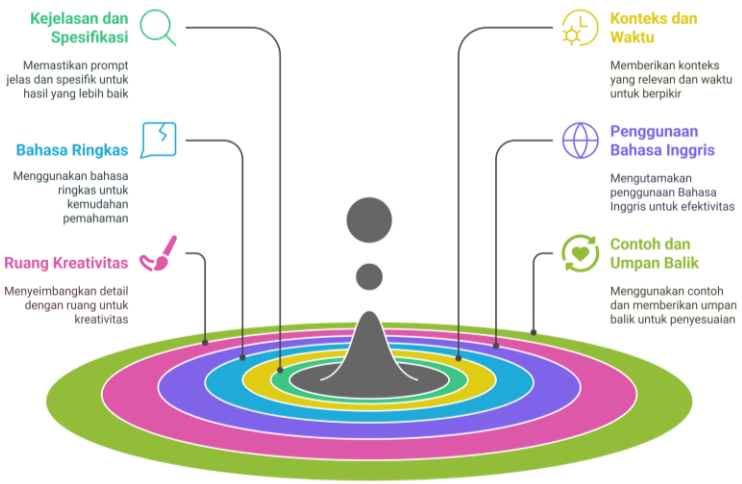


1-3 Prompting yang kompleks

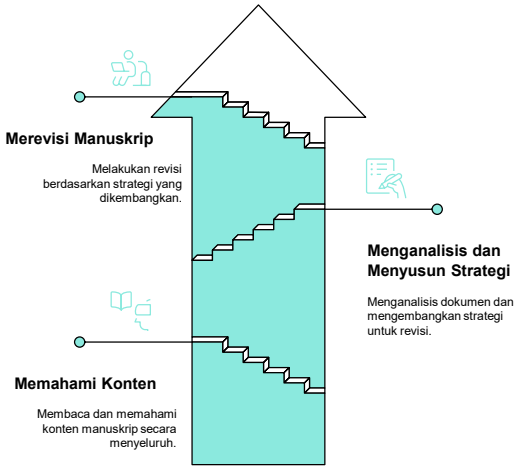
## Teknik Prompting: Retrieval-Augmented Generation



# Teknik Prompting: Instruction-Based Prompting
## AI → *Organisme Digital*

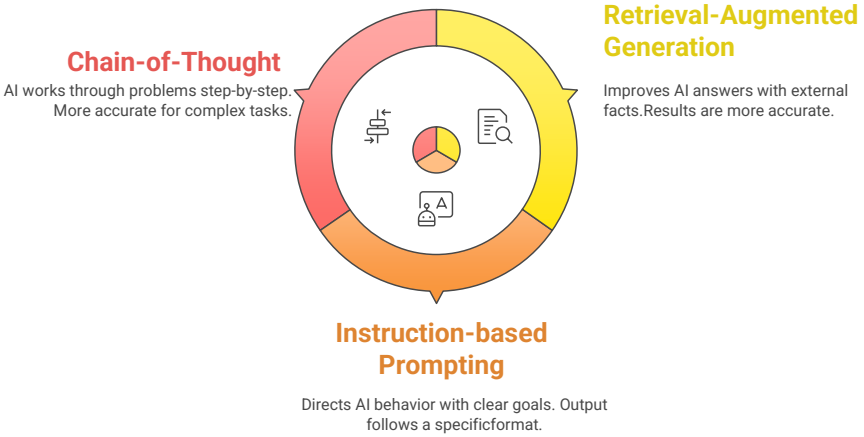



# Teknik Prompting: *Chain-of-Thought*

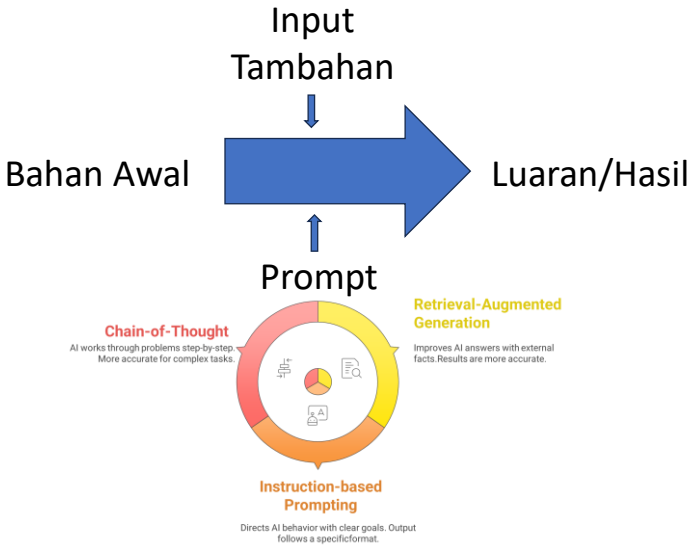


# Teknik Prompting Lanjut

**Pilihan Model**
**Tipe respon: Chat atau Canvas**




# Proses Prompting

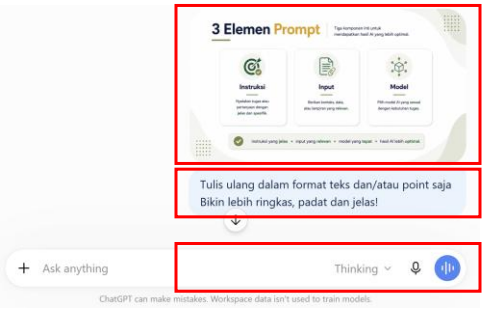


# 3 Elemen Prompt

**Instruksi:** Tugas atau pertanyaan.

**Input** : Konteks, data, atau lampiran.

**Model** : Pilih model AI.




**Input**

**Prompt**

**Model**

**Rumus sederhana:**
Instruksi jelas + input relevan + model tepat = hasil AI lebih optimal.

# Pilihan Prompting

Pilih metode prompting sesuai kebutuhan dan tujuan kerja Anda.

**Standard — Prompt Detail/Panjang**

- **Panjang/Susah**: Prompting detail, harus **menguasai materi** dan **berpengalaman**
- **Liar**: Variasi respond tinggi, memerlukan revisi berkali-kali
- **Fleksibel**: Prompt bisa dirubah sesuai kemauan

**Custom-GPT — Shortcut Prompt (Prompt detail di dalam sistem)**

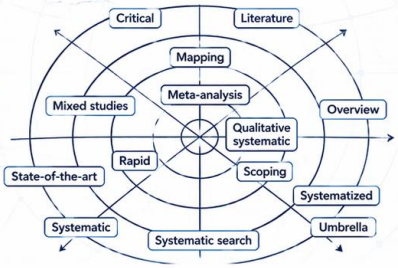
- **Singkat**: prompt pendek
- **Cepat**: hasil instan untuk kebutuhan khusus (i.e., Menulis Jurnal)
- **Terbatas**: perlu mendapatkan prompt dari developer
- **Rigid**: terbatas untuk fungsi tertentu

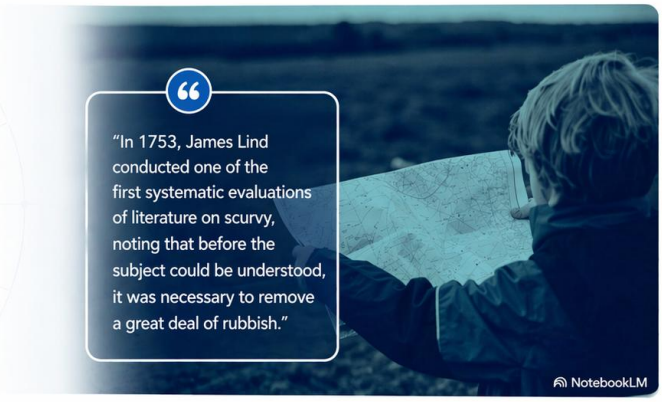
# Dari Lautan Informasi Menuju Sintesis Pengetahuan

Tantangan utama penelitian: menavigasi beragam metodologi tinjauan.

**14 JENIS TINJAUAN**

Peneliti harus menavigasi 14 jenis tinjauan yang berbeda.




## MESIN UNIVERSAL SINTESIS LITERATUR

**1 Search**
Identifying the literature.
- Exhaustive vs. Selective
- Time-bound vs. Open

**2 Appraisal**
Evaluating quality and bias.
- Standardized checklists
- Informal critique
- None

**3 Synthesis**
Combining the findings.
- Narrative/Thematic
- Tabular
- Statistical/Meta-analysis

**4 Analysis**
Interpreting the data.
- Chronological
- Conceptual
- Effect-size measurement

**Setiap jenis tinjauan adalah konfigurasi unik dari input dan batasan SALSA.**

# 10 Tahap Sistematis, Insight Bukan Kebetulan.

Satu langkah terlewat, seluruh hasil bisa bias.


**Key Insight:**
Systematic review bekerja seperti rantai. Kuat hanya sekuat mata rantai terlemahnya.

**Tidak ada shortcut dalam ilmu.** Ikuti semua tahap, hasil akan berbicara.

# Perbedaan Empat Jenis Tinjauan

Narrative Review • Scoping Review • Systematic Review • Meta-Analysis

| | 1. NARRATIVE REVIEW | 2. SCOPING REVIEW | 3. SYSTEMATIC REVIEW | 4. META-ANALYSIS |
|---|---|---|---|---|
| TUJUAN | Pemahaman luas topik melalui sintesis naratif. | Memetakan bukti: topik, konsep, jenis bukti, celah. | Menjawab pertanyaan spesifik dengan sintesis sistematis. | Menggabungkan hasil numerik dari studi serupa. |
| CAKUPAN | Luas, tanpa batasan ketat metode/desain. | Luas, berbagai jenis sumber & desain. | Spesifik, fokus pada pertanyaan jelas. | Spesifik, studi kuantitatif dengan kriteria ketat. |
| METODE | Fleksibel, tidak selalu prosedur baku. | Scoping (mis. PRISMA-ScR) iteratif. | Protokol ketat (mis. PRISMA) sistematis & replikatif. | Teknik statistik untuk menggabungkan *effect size*. |
| HASIL | Sintesis naratif, interpretasi, perspektif konseptual. | Peta bukti: tema, konsep, jenis intervensi, populasi, celah. | Sintesis tematik/kualitatif atau naratif untuk menjawab pertanyaan. | Estimasi efek gabungan (*effect size*) (mis. CI, *p*-value). |
| KEGUNAAN | Pemahaman awal, identifikasi konsep & arah penelitian. | Identifikasi cakupan, tren, dan celah penelitian. | Dasar keputusan berbasis bukti (praktik, kebijakan, penelitian). | Menentukan besaran efek intervensi/variabel secara lebih akurat. |
| KARAKTERISTIK KUNCI | Subjektif, berfokus pada interpretasi penulis. | Eksploratif, inklusif, memetakan lanskap bukti. | Sistematis, transparan, dapat direplikasi. | Kuantitatif, berbasis statistik, homogenitas studi diperlukan. |


Claim of Novelty

Research gap – Novelty Hypothesis – Objectives/Research questions

Data: Proof – Evidence

Confirmation of the claim
Answers to the research questions

Title

Abstract, Keyword

Introduction

Materials and Methods

Results Discussion

Conclusion

References

**Dalam SLR, Literatur → Data**

Data set 1

Data set 2

Data set 3

Data set 4

Title

Abstract

Introduction

Methods
**(Structured, Systematic)**

Results and Discussion

Data set 1

Data set 2

Data set 3

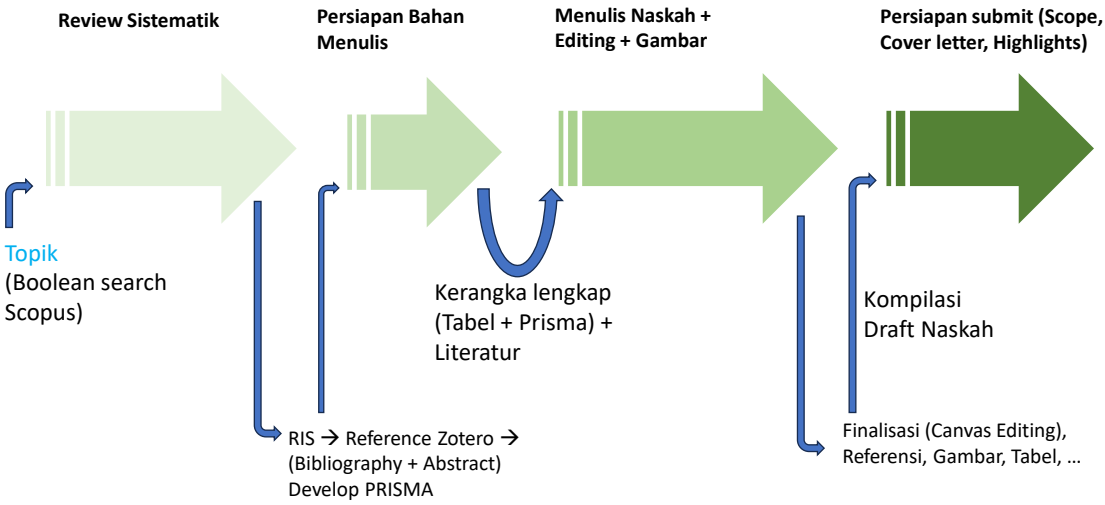
Data set 4

Conclusion

References

**Progres Penelitian**

**Workflow**: SLR with GRA




## Standar Kerangka SLR

**Menulis: GPT 5 (Auto, FAST, Thinking dan Scite.AI**

- Title → **Scopus + ChatGPT**
- Abstract
1. Introduction
2. Methodology (Figure 1: PRISMA) → **Scopus/WoS/PubMed + ChatGPT**
3. Fundamental Concept (Figure 2)

4. Important Findings
   - 4.1 Theme 1 (Table 1) (Figure 3)
   - 4.2 Theme 2 (Table 2) (Figure 4)
   - 4.3 Theme 3 (Table 3) (Figure 5)
   - 4.4 Theme 4 (Table 4) (Figure 6)

→ **Ekstraksi Data ChatGPT**

**Finalisasi:**
- Rapikan Tabel
- Integrasikan gambar (Letakkan pada paragraf kedua untuk setiap sub-bab)

5. Discussion
6. Conclusion and Perspectives

References → **Scopus + ChatGPT**

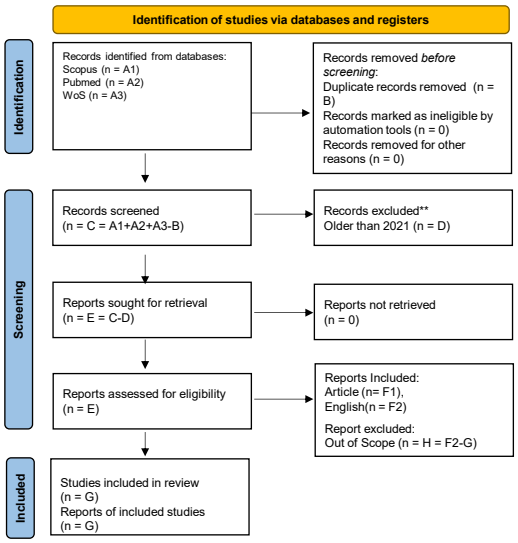
Visualisasi → **GPT + Powerpoint**

## Prompting Framework SLR

| Input/Model | Prompt | Output + action |
|---|---|---|
| Topik riset<br>Instant | **Prompt 0**: Suggest Boolean Search for research topic! [Topic: ……]! | Keywords pencarian literatur Scopus (hanya satu database).<br>Lakukan pencarian (Target 100-500), buat file References_Summary<br>Mulai buat diagram Prisma! |
| References_Summary_SLR<br>Thinking | **Prompt 1**: Suggest titles! | Tiga topik pilihan (pindahkan ke file MS Word, validasi, pilih sesuai minat) |
| Thinkig | **Prompt 2**: Write the outline! | Canvas, kerangka keseluruhan (gabungkan ke file MS Word) |
| Thinking | **Prompt 3**: Extract data to construct Tables! | Canvas, Table 1-4, gabungkan ke file MS Word.<br>Selesaikan Prisma, gabungkan ke file MS Word. |
| Thinking | **Prompt 4**: Generate references statements | Pernyataan literatur, gabungkan ke file MS Word.<br>Boleh mengambil dari references summary saja (dengan prompt tambahan) |
| File **MS Word Extended** | **Prompt 5**: Write SLR manuscript | File Draft #1 SLR |
| Thinking | **Prompt 6**: Propose figures! | Deskripsi, judul dan penjelasan gambar, gabungkan ke file word |
| Instant | **Prompt 7**: Draw Figure XX! | Gambar (2-6) satu per satu, gabungkan ke file Draft #1 SLR |
| Instant | **Prompt 8:** Write captions and description! | Keterangan gambar XX (2-6) dan deskripsi, gabungkan ke file Draft #1 SLR |

Catatan: Untuk menghitung jumlah referensi unik (included) di Table 1-4, gunakan prompt tambahan.

## Diagram Prisma (Custom-GPT GRA)



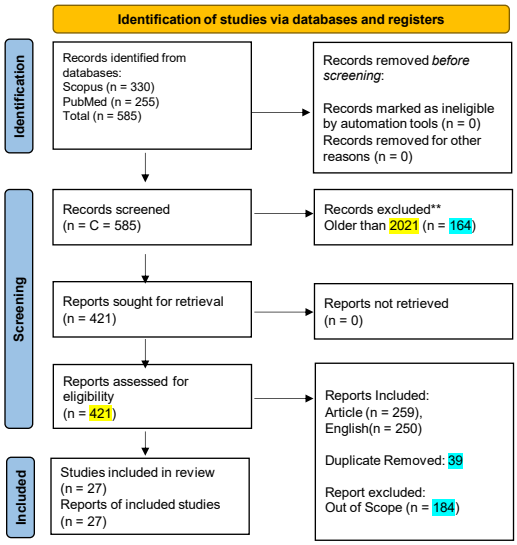

### Kelebihan

- Cepat
- AI membantu dongkrak kualitas
- Cocok untuk bidang riset non-sensitive (control longgar)

### Kekurangan

- Satu database (scopus)
- Exclusi dalam database suharusnya bukan bagian integral PRISMA
- Berbasis abstrak saja (tidak sesuai protocol)
- Tidak sesuai bidang riset
- Tidak terdaftar (i.e., Prospero)
- Tidak ada analisis Risiko bias (*Risk of Bias*)
- Tidak ada proses internal review (>1)

## Diagram Prisma (Custom-GPT GRA) – Scopus + PubMed



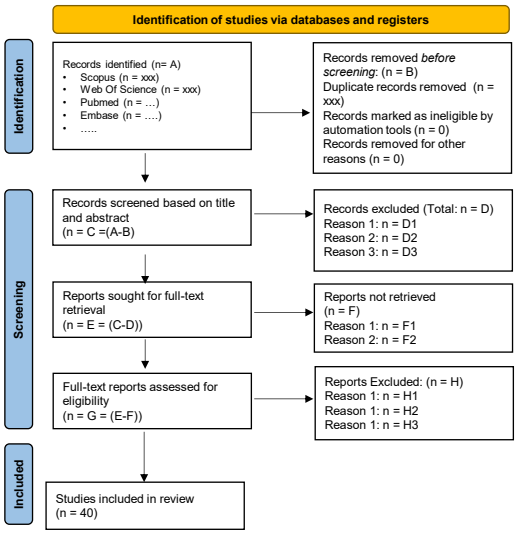

( ( "cancer patient*" OR "oncology patient*" OR "patients with cancer" OR neoplasm* ) AND ( "malnutrition" OR "sarcopenia" OR "cancer cachexia" ) AND ( "prediction model*" OR "risk prediction model*" OR "clinical prediction rule*" OR "risk score*" OR nomogram* ) )

## Integrasi: Rayyan + Custom-GPT GRA)

| | | |
|---|---|---|
| Topik → Judul → Proposal/Kerangka | → | GPT + Databases |
| Search String (Multiple Database) | → | GPT + Databases |
| Import Data | → | RAYYAN + Databases |
| Remove Duplicate | → | RAYYAN |
| Screening: Abs+Title | → | Export RAYYAN → Screen GPT<br>Update RAYYAN → Download Full-text |
| Screening: Full-text | → | Full text (Input RAYYAN) → Screen GPT<br>Update RAYYAN → Generate **Prisma** |
| Included: Full-text | → | Full text → **Extract GPT** → **Writing** |

→ **Prospero**

## Diagram Prisma (RAYYAN)

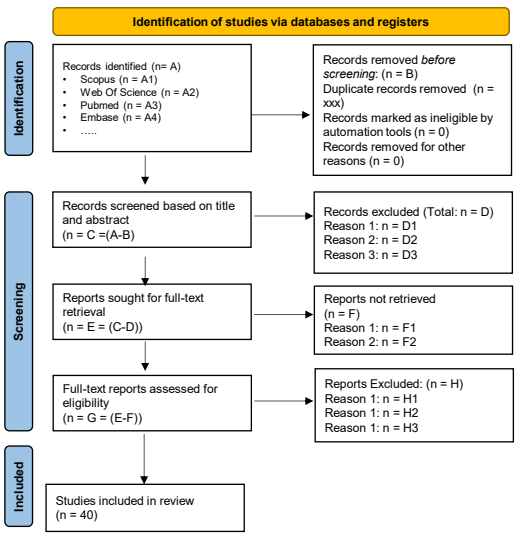



**Workflow RAYYAN**
- Import Data (Multi-databases)
- Remove Duplicate (Otomatis)
- Screening Abs+Title: (Manual, >1 Peneliti)
- Screening Full-text : (Download → Upload → Pilih)
- Included: Full-text

**Kelebihan**
- Proses Transparan
- Prudent (>1 peneliti)
- Record keeping
- Bagian metodologi → Supplementary

**Kelemahan**
- (sangat) Lama
- Hilang momentum dan focus
- Hasil terbatas keahlian peneliti

## Integrasi: Rayyan + Custom-GPT GRA)




Identifikasi study dengan kata kunci (**Prompt 0**)
- Database 1 (n=A1)
- Database 2 (n=A2),
- Database 3 (n=A3)
- …….
- (Total, n=A = (A1 + A2 + A3+ …….)

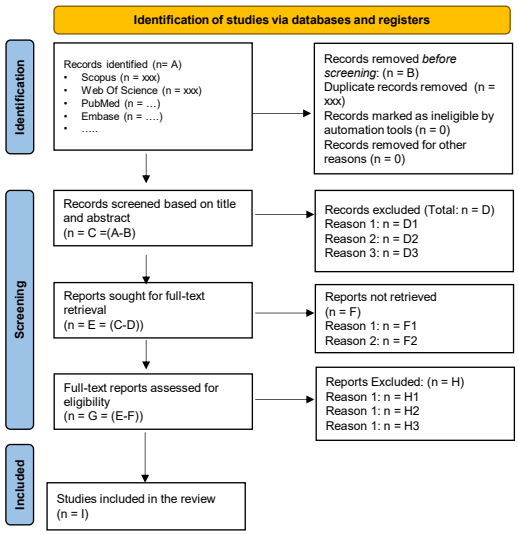
Record screened: n = C = (A-B)

Import semua ke RAYYAN, deduplikasi (n = B)

Export Bibtex → Import Zotero
→ Buat References Summary

**Prompt 1 → Pilih topik sesuai minat dan keahlian**
**Prompt 2 → Buat kerangka awal → Ajukan ke Prospero**

## Integrasi: Rayyan + Custom-GPT GRA)

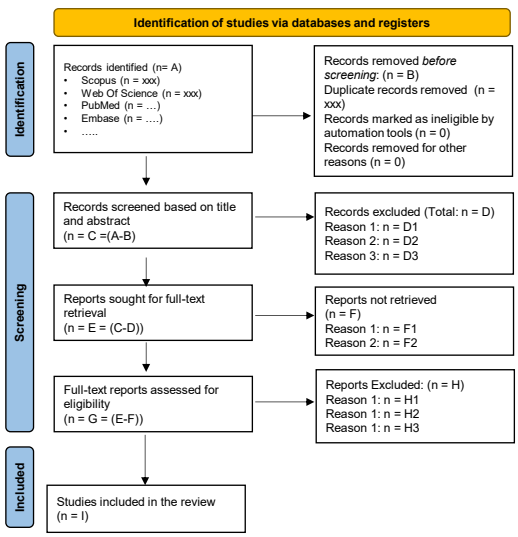



Record screened: n = C = (A-B)
Proses Screening → Export Bibtex → Import Zotero → Buat References Summary
**Prompt 3** → (Tambahkan alasan eksklusi, kelompokkan jadi 3-4 kategori**)**

Record Excluded, Total: n = D
Reason 1: D1
Reason 2: D2
Reason 3: D3

Download Full text semua artikel terpilih included) (Total jumlah artikel = E)
Artikel yang tidak diperoleh (Total F), Alasan:
Reason 1: F1
Reason 2: F2
Reason 3: F3
Full-text reports assessed for eligibility, (n = G = (E-F))

## Integrasi: Rayyan + Custom-GPT GRA




**[Lampirkan semua PDF n = G]**
**Prompt 3 (Ulang)** → Refine Tables 1-4 based on full text. Remove studies out of scope (if any) found after assessing full-text! Add exclusion reasons (if any) grouped into 3-4 categories (if possible).

Reports Excluded: (n = H)
Reason 1: n = H1
Reason 1: n = H2
Reason 1: n = H3

Studies included in the review
(n = I) → Seluruh artikel full text pada Tabel 1-4.

Lanjutkan: **Prompt 4, Prompt 5, Prompt 6, Prompt 7, Prompt 8!**

## Prompting Framework: Integrasi: Rayyan + Custom-GPT GRA

| Model | Input | Prompt | Output + action |
|---|---|---|---|
| 5.5 Instant | Topik riset | **Prompt 0**: Suggest Boolean Search for research topic [Topic: ……] in several suitable databases (provide the 3-5 suitable list! | Keywords pencarian literatur Scopus (**gunakan multi-database**).<br>Lakukan pencarian (**Target 100-500 untuk masing-masing database**), buat file References_Summary<br>Mulai buat diagram Prisma! (**Gunakan template RAYYAN**)<br>Lakukan deduplikasi di RAYYAN<br>Data Prisma: A (AB, A2, A3, …..), B (B1, B2, B3, …), C = A-B |
| 5.5 Thinking | References_Summary_SLR | **Prompt 1**: Suggest titles! | Tiga topik pilihan (pindahkan ke file MS Word, validasi, pilih sesuai minat) |
| 5.5 Thinking | | Title:<br>**Prompt 2**: Write the outline! Prepare in an extensive form (3000-5000 words) suitable for Prospero reporting! Respond in Chat! | Canvas, kerangka keseluruhan (gabungkan ke file MS Word) |
| 5.5 Thinking | Full teks semua artikel inklusi | **Prompt 3**: Extract data to construct Tables! Aim for 40-60 most relevant articles! (Tambahkan alasan eksklusi, kelompokkan alasan menjadi jadi 3-4 kategori**)** | Table 1-4 (Sementara).<br>Download full-teks semua study inklusi<br>Update Data Prisma: D (Alasan eksklusi: D1, D2, D3, ….),<br>Data n = E = C-D<br>Data F untuk yang tidak berhasil di download (F1, F2, F3,…)<br>Data G = E-F |

## Prompting Framework: Integrasi: Rayyan + Custom-GPT GRA

| Model | Input | Prompt | Output + action |
|---|---|---|---|
| | Semua full-text | **Prompt 3**: Refine Tables 1-4 based on full text. Remove studies out of scope (if any) found after assessing full-text! Add exclusion reasons (if any) grouped into 3-4 categories (if possible). | Table 1-4 (Final), gabungkan ke file MS Word.<br>Update Data Prisma: F (Alasan eksklusi: H1, H2, H3, ….)<br>Data I = G-F |
| 5.5 Thinking | | **Prompt 4**: Generate references statements | Pernyataan literatur, gabungkan ke file MS Word.<br>Boleh mengambil dari references summary saja (dengan prompt tambahan)<br>Selesaikan Prisma, gabungkan ke file MS Word. |
| 5.5 Thinking **Extended** | File **MS Word** | **Prompt 5**: Write SLR manuscript | File Draft #1 SLR |
| 5.5 Thinking | | **Prompt 6**: Propose figures! | Deskripsi, judul dan penjelasan gambar, gabungkan ke file word |
| 5.5 Thinking | | **Prompt 7**: Draw all Figures! | Gambar, gabungkan ke file Draft #1 SLR |
| 5.5 Thinking | | **Prompt 8:** Write captions and description! | Keterangan gambar dan deskripsi, gabungkan ke file Draft #1 SLR |