

大規模言語モデルと 言語資源

国立情報学研究所

相澤彰子

本日の話題

- 大規模言語モデルの構築に使われる言語コーパスとは
- 大規模言語モデルが生成するテキストの特徴

大規模言語モデルの構築に使 われる言語コーパスとは

<https://life architect.ai/models-table/>による
言語モデル近況

Model	Lab	Playgroun d	Parameter s (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling≥ 20:1)	ALScore "ALSore 20:1" is a quick and dirty	MML U	MML U -Pro	GPQA	HLE	Training dataset	Announce d ▼	Public? ●	Paper / Repo	Arch	Tags
Step 3	StepFun	https://www.s	321	18000	57	8.0			72.9		web-scale	Jul/2025	●	https://git	MoE	
Qwen3-235B-A22B-	Alibaba	https://hugging	235	36000	154	0.7	02.8	84.4	81.1		synthetic, web-scale	Jul/2025	●	https://hu	MoE	Reasoning
KAT-V1-200B	Kuaishou		200	10000									●	https://ar	MoE	Reasoning
KAT-V1-40B	Kuaishou	https://hugging	40	10000									●	https://ar	Dense	Reasoning
Qwen3-Coder-480B-	Alibaba	https://hugging	480	36000									●	https://qv	MoE	SOTA
Qwen3-235B-A22B-	Alibaba	https://hugging	235	36000									●	https://hu	MoE	
FlexOlmo	Allen AI	https://hugging	37	4150									●	https://ar	MoE	
EXAONE 4.0	LG	https://hugging	32	14000									●	https://w	Dense	Reasoning
Kimi K2	Moonshot AI	https://hugging	1000	15500									●	https://m	MoE	Reasoning, SOTA
Reka Flash 3.1	Reka AI	https://hugging	21	5000									●	https://w	Dense	Reasoning
Devstral Medium	Mistral	https://chat.m	50	12000	240	2.0					synthetic, web-scale	Jul/2025	●	https://m	Dense	
Grok 4	xAI	https://grok.co	5000	80000	16	66.7			88.9	44.4	synthetic, web-scale	Jul/2025	●	https://lif	MoE	Reasoning, SOTA
Phi-4-mini-flash-reas	Microsoft	https://hugging	3.8	5150	1356	0.5					synthetic, web-scale	Jul/2025	●	https://az	Dense	
T5Gemma		https://hugging	9	10000	1112	1.0	76.7	55.7	40.4		web-scale	Jul/2025	●	https://de	Dense	
MedGemma	Google DeepMin	https://hugging	27	14000	519	2.0	87				web-scale	Jul/2025	●	https://arx	Dense	
R1T2 Chimera	TNG	https://hugging	685	14800	22	10.6					synthetic, web-scale	Jul/2025	●	https://ar	MoE	
Spectra 1.1	Consortium		3.6	1200	334	0.2	36.1				synthetic, web-scale	Jun/2025	●	https://ar	Dense	
DiffuCoder	Apple	https://github	7	5630	805	0.7					code, The Stack	Jun/2025	●	https://ar	Dense	Diffusion
Hunyuan-A13B	Tencent	https://hugging	80	7000	88	2.5	88.2	67.2	71.2		synthetic, web-scale	Jun/2025	●	https://hu	MoE	
Mercury	Inception	https://chat.in	90	8000	89	2.8		69	51	3.4	synthetic, web-scale	Jun/2025	●	https://w	Dense	Diffusion
Mu	Microsoft	https://blogs.w	0.5	500	1000	0.1					synthetic, web-scale	Jun/2025	●	https://bl	Dense	
Gemini Robotics On-	Google DeepMin	https://docs.gc	20	10000	500	1.5					synthetic, web-scale	Jun/2025	●	https://de	MoE	
ICONN-1	ICONNAI	https://hugging	88	10000	114	3.1					synthetic, web-scale	Jun/2025	●		MoE	

FineWeb コーパス構築の例

データセット公開サイト：

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

“15 trillion tokens of the finest data the  web has to offer”

The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf
NeurIPS 2024 Track Datasets and Benchmarks Spotlight
<https://arxiv.org/abs/2406.17557>

FineWeb2: One Pipeline to Scale Them All -- Adapting Pre-Training Data Processing to Every Language

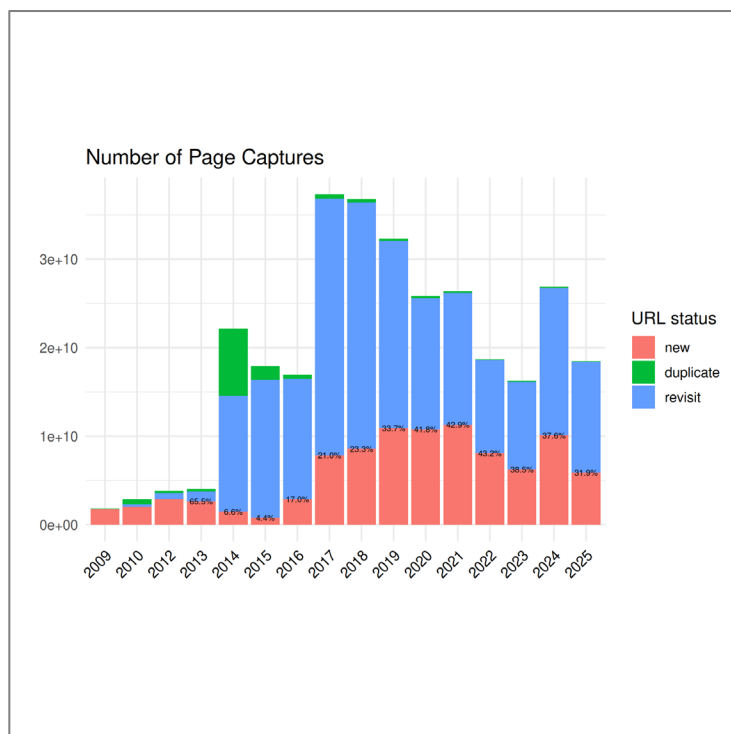
Penedo, Guilherme, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025.
arXiv [Cs.CL]. arXiv. <http://arxiv.org/abs/2506.20920>.

15 Trillion トークン

- 150000000000000 トークン
 - 2 億冊の本に相当
 - 国会図書館12館分
 - 恐らく、これまで出版された本より多い量のテキスト
-
- 参考：BCCWJ 1億語（短単位）

Common Crawl <https://commoncrawl.org/>

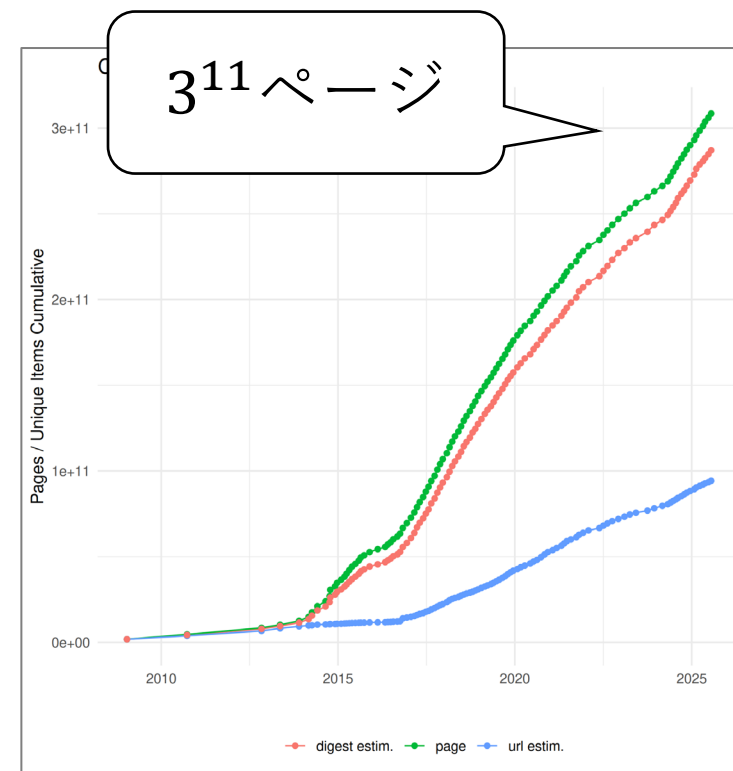
- Common Crawlは、研究および分析の目的でインターネット（Web文書）のコピーを無償で提供する非営利団体



◀年ごとの収集データ

データの累積サイズ▶
(ページ)

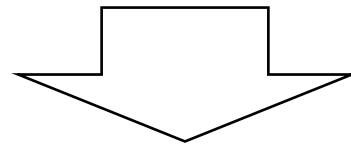
<https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize>



FineWeb

“The performance of a large language model (LLM) depends heavily on the **quality** and size of its pretraining dataset.”

“大規模言語モデル（LLM）の性能は、その事前学習データセットの**品質**とサイズに大きく依存する。”



「品質」をどうやって向上させるか？

Web ページ $\neq$ テキスト

```
<!DOCTYPE html><!-- Last Published: Fri Jul 25 2025 15:59:28 GMT+0000 (Coordinated Universal Time) --><html data-wf-domain="commoncrawl.org" data-wf-...
<style>
  /*
  -webkit-font-smoothing: antialiased;
  -moz-osx-font-smoothing: grayscale;
  */
  /* Invert the scrollbar colors */
  ::-webkit-scrollbar {
    width: 12px;
    background-color: white; /* Lighter shade */
  }
  /* Invert the scrollbar thumb colors */
  ::-webkit-scrollbar-thumb {
    background-color: #d1d5db; /* Slightly darker shade */
    border-radius: 6px;
  }
  /* Invert the scrollbar thumb colors on hover */
  ::-webkit-scrollbar-thumb:hover {
    background-color: #e2e8f0; /* Slightly lighter shade on hover */
  }
  /* Invert the scrollbar thumb colors when active (being clicked/dragged) */
  ::-webkit-scrollbar-thumb:active {
    background-color: #c3c9d1; /* Lighter shade when active */
  }
</style>
<link rel="help" href="https://index.commoncrawl.org/ccbot.json">
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "FAQPage",
  "mainEntity": [
    {
      "@type": "Question",
      "name": "What is Common Crawl?",
      "acceptedAnswer": {
        "@type": "Answer",
        "text": "Common Crawl is a 501(c)(3) non-profit organization dedicated to providing a copy of the Internet to Internet researchers, co
      }
    },
    {
      "@type": "Question",
      "name": "What can you do with Common Crawl data?",
      "acceptedAnswer": {
        "@type": "Answer",
        "text": "The possibilities are endless. People have used the data to improve language translation software, predict trends, track dis
      }
    },
    {
      "@type": "Question",
      "name": "Can't Google or Microsoft just do what Common Crawl does?",
      "acceptedAnswer": {
        "@type": "Answer",
        "text": "Our goal is to democratize the data so that everyone, not just big companies, can do high-quality research and analysis."
      }
    },
    {
      "@type": "Question",
      "name": "Under what terms is Common Crawl data released?",
      "acceptedAnswer": {
        "@type": "Answer",
        "text": "As strong believers in Open Data, we apply as few restrictions as possible to the dataset.
        The terms we add (primarily in an effort to prevent abusive or illegal usage) are described on our Terms of Use page."
      }
    }
  ]
}
```

◀ ページソース

クリーニング

英語以外の文書
定型文
メニューテキスト

...

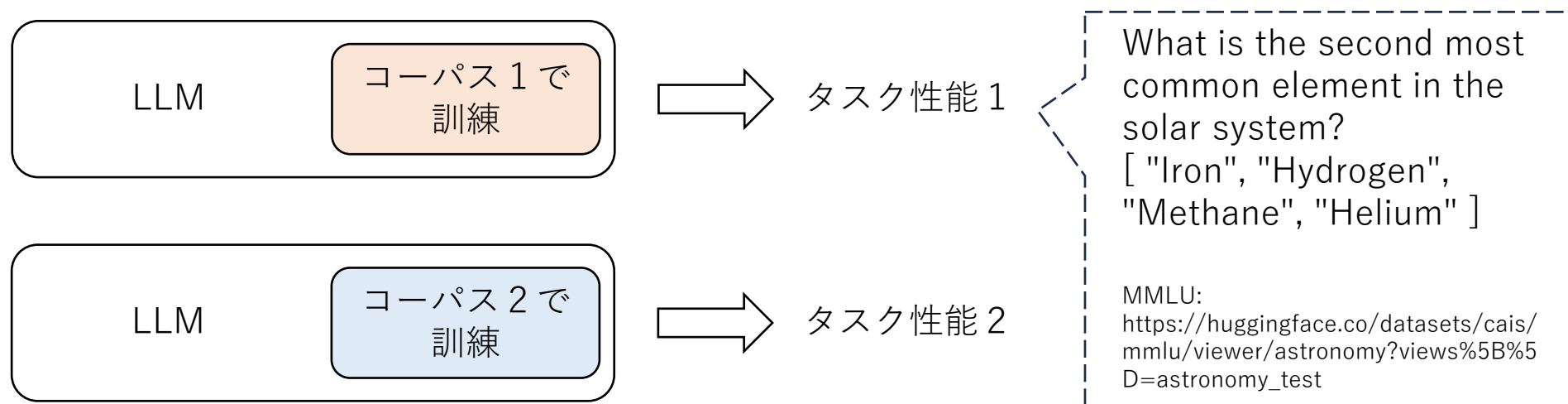
抽出される“テキスト”

*What is Common Crawl?
Common Crawl is a
501(c)(3) non-profit
organization dedicated to
providing a copy of the
Internet to Internet
researchers, companies
and individuals at no cost
for the purpose of research
and analysis.*

....

コーパスの「品質」の測り方

- サンプルしたコーパスで小さなLLMを訓練する
- 自然言語処理のベンチマーク問題を解かせて性能を測定



- 高い性能が得られる方が「よい」コーパス

コーパスのクリーニング

- ウェブページからテキストを抽出
- 「同じ」とみなされる文書が複数ある場合、重複を除く
- 経験則に基づくフィルターを適用

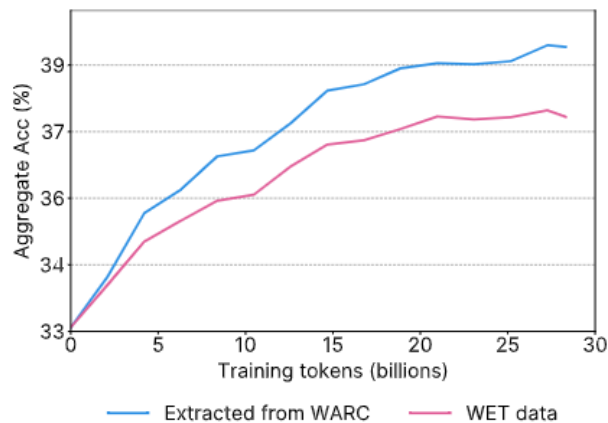


Figure 1: **Trafilatura-extracted WARC vs WET** 28B tokens ablation study. Custom text extraction outperforms the default WET data. No filtering or deduplication was applied except fastText English language filtering.

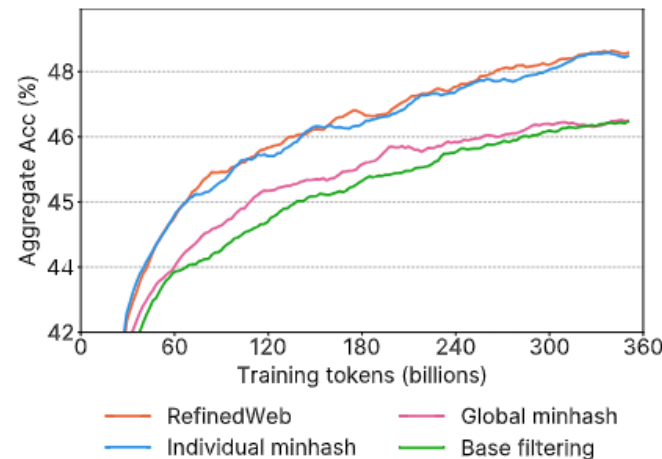


Figure 5: **Individual minhash deduplication study.** Unlike Global minhash, deduplicating individually improves the average score.

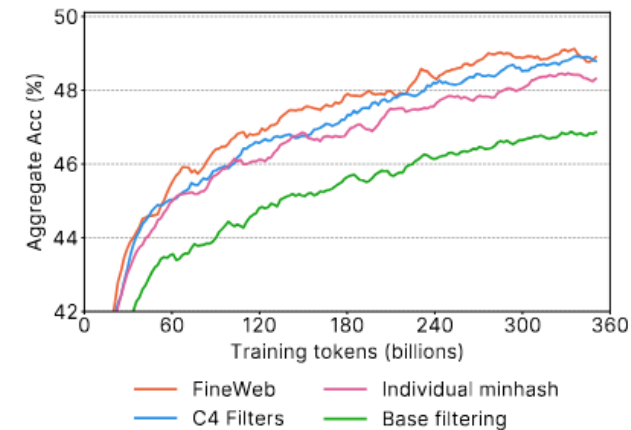


Figure 9: **Each processing step in FineWeb provides a performance uplift.** Compared to the base filtering (Section 3.3), applying individual-crawl Min-Hash deduplication (Section 3.4) the C4 filters (Section 3.5), and our additional heuristic filters (Section 3.6) each improve performance.

前処理の
積み重ね
で少しずつ
性能改善

さらなる品質管理：FineWeb-Edu

- 15兆トークンのコーパスから「教育的価値（educational value）」の高い文書だけを集めて 1.3兆トークンのサブコーパスを構築
- 手順
 - 既存のLLM（Llama-3-70B-Instruct）を使って、サンプルした460,000ウェブページの「教育的価値」を0～5 スケールで自動評価
 - その結果を正解だとみなして軽量の自動分類器を訓練
 - 分類器をコーパスの全文書に適用

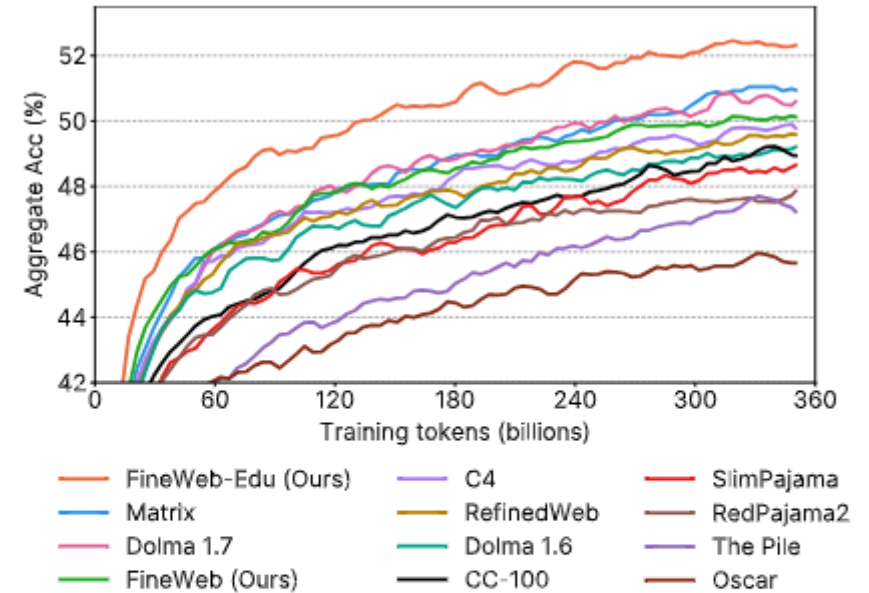


Figure 10: Comparing FineWeb datasets to other public datasets. Base FineWeb shows strong performance, with the educational subset (FineWeb-Edu) surpassing all other public datasets and further enhancing the aggregate score by approximately 2%.

教育的価値を判定するためのプロンプト

以下はウェブページからの抜粋です。以下の5段階評価システムを使用して、このページが小学校から中学校までの教育現場で教育的な価値が高く、教育に役立つかどうかを評価してください。各基準を満たすごとにポイントが加算されます：

- 抜粋が教育的なテーマに関連する基本的な情報を提供している場合（広告や宣伝資料などの関連性のない内容や非学術的な内容を含む場合でも）、1点を加算します。
- 抜粋が教育に関連する特定の要素を扱っているが、教育基準と密接に一致していない場合、さらに1点を加算します。教育コンテンツと非教育コンテンツを混合し、有用なトピックの表面的な概要を提供したり、情報が整理されていない、または論理的でない文章スタイルで提示されている場合があります。
- 抽出物が教育目的に適しており、学校カリキュラムに関連する重要な概念を導入している場合、3点を付与します。文章は一貫性がありますが、包括的ではない可能性があります。余分な情報を含む場合があります。教科書や基本的なチュートリアルを導入部分のようなもので、学習に適していますが、小学校の生徒には複雑すぎる概念を扱うなどの顕著な制限があります。
- 4点を与える場合、抽出物が小学校レベル以下の教育目的に極めて関連性が高く有益であり、明確で一貫した文章スタイルを示している場合。教科書の一章やチュートリアルに類似し、演習問題と解答を含む実質的な教育コンテンツを提供し、不要な情報は最小限で、小学校の生徒にとって概念が過度に高度でない場合。内容は一貫性があり、焦点を絞り、構造化された学習に価値がある。
- 抽出物が教育的な価値において卓越しており、小学校または小学校レベルでの教育に完全に適している場合、5点を与えます。詳細な論理展開を伴い、文章スタイルは理解しやすく、テーマに関する深い洞察を提供し、非教育的な内容や複雑な要素を含まないもの。

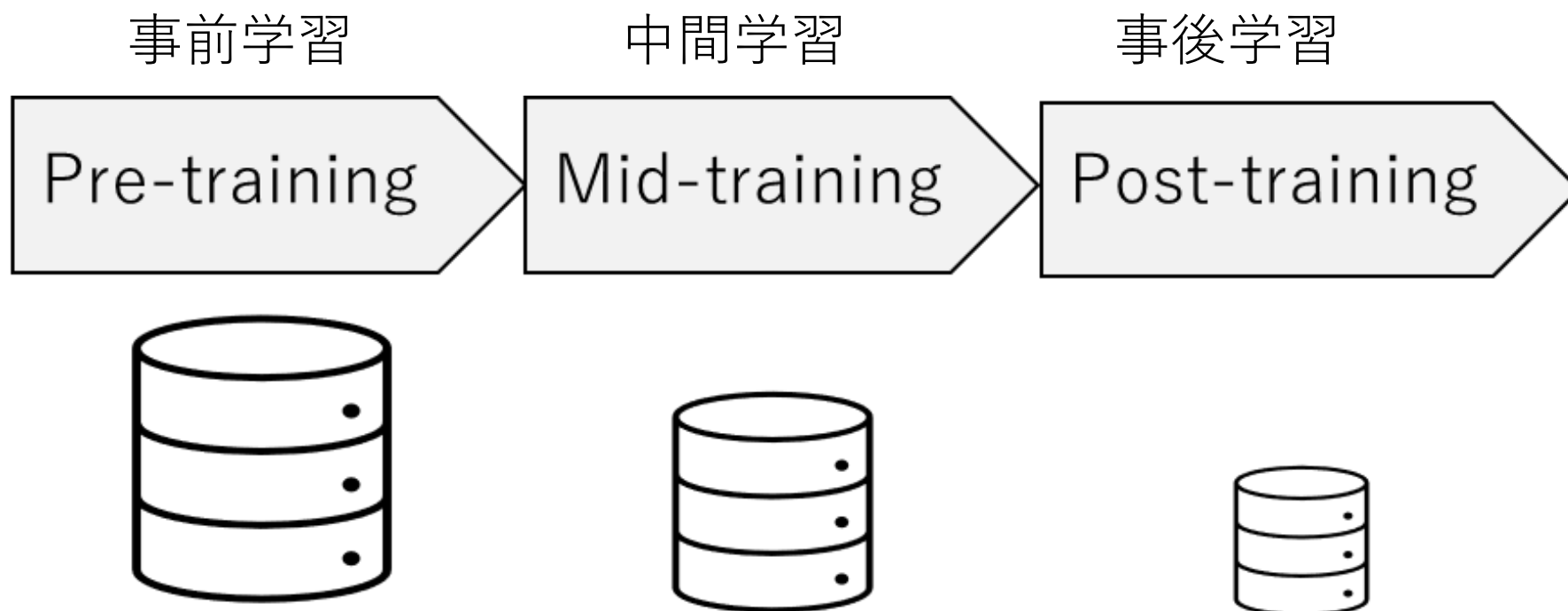
抜粋: <EXAMPLE>

抜粋を検討した後に：

- 総得点を100字以内で簡潔に説明して下さい。
- 最後に “Educational score: <total points>” というフォーマットでスコアを示して下さい。

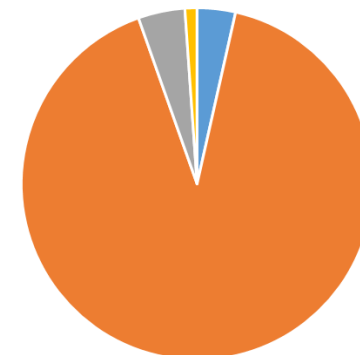
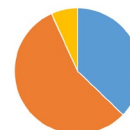
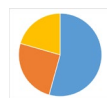
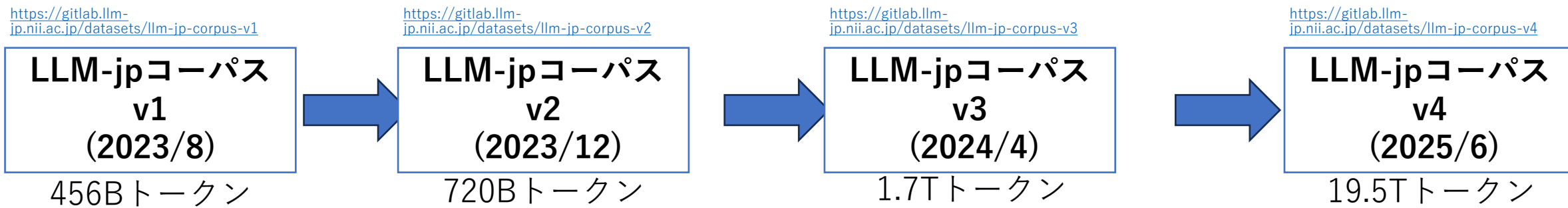
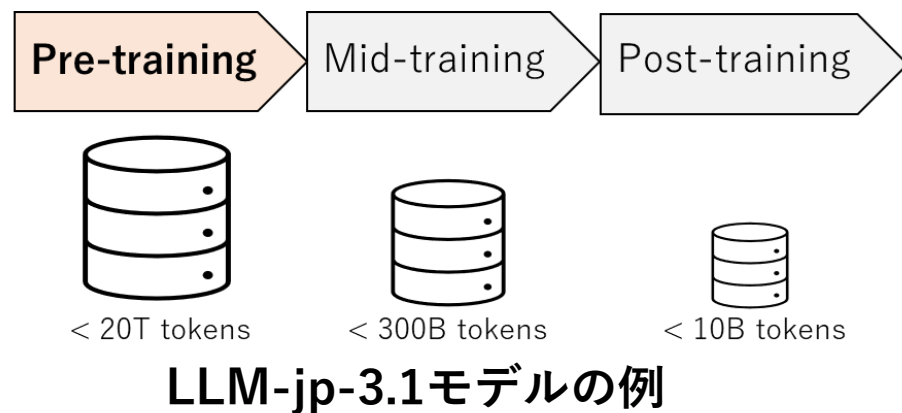
DeepLの翻訳を人手で確認

言語モデルの学習



事前学習 (Pre-Training)

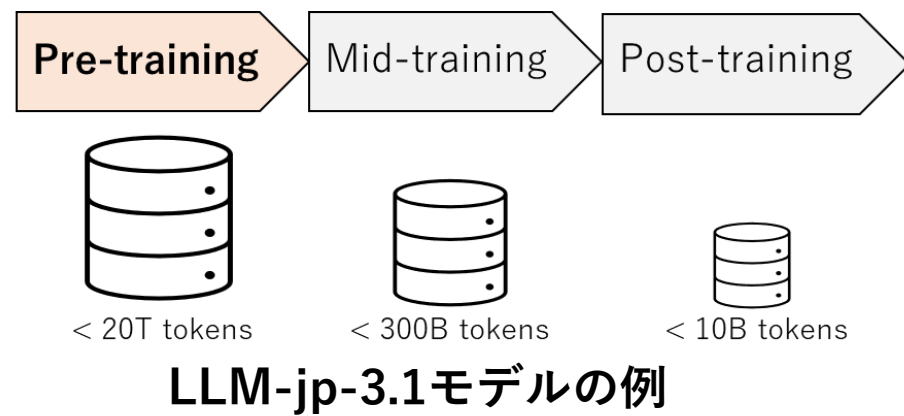
- 大きなコーパスを使った大規模なモデル学習
- コーパスは目的に応じて配合



■ 日本語 ■ 英語 ■ 中国語と韓国語 ■ コード*

事前学習 (Pre-Training)

- 大きなコーパスを使った大規模なモデル学習
- コーパスは目的に応じて配合



<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v1>

LLM-jpコーパス
v1
(2023/8)

456Bトークン

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	mC4	136B
英語	Wikipedia	5B
	Pile	176B
コード	Stack	148B

<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v2>

LLM-jpコーパス
v2
(2023/12)

720Bトークン

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	CC	390B
英語	Wikipedia	5B
	Pile	176B
コード	Stack	148B

<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3>

LLM-jpコーパス
v3
(2024/4)

1.7Tトークン

言語	サブコーパス	トークン数
日本語	Wikipedia	1B
	CC	380B
	NDL PDF/HTML	207B
英語	Wikipedia	5B
	Dolma	945B
他言語	中韓Wikipedia	1B
コード	Stack	114B

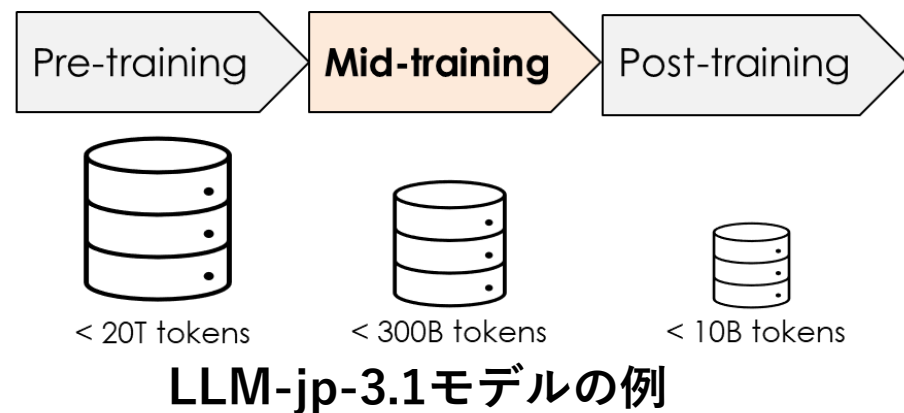
<https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v4>

LLM-jpコーパス
v4
(2025/6)

19.5Tトークン

言語	サブコーパス	トークン数
日本語	FineWeb-2	236B
	CC	223B
	Web PDF/HTML	100B
	Others	129B
英語	FineWeb	17.6T
	Others	200B
他言語	中韓	841B
コード	Stack, StarCoder	218B

中間学習 (Mid-Training)



- 中間学習: 事前学習と事後学習の間に行う, 中規模のモデル学習
- たとえば下流タスクに近い指示応答ペアを合成

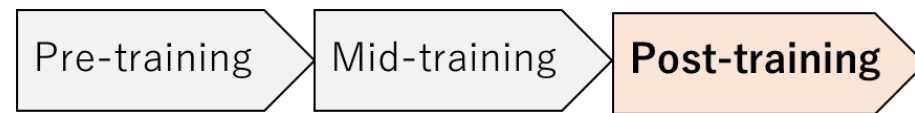
Instruction:

『リンクの冒険』に関する文章から、ゲームの発売日と発売プラットフォームを抽出してください。

『リンクの冒険』（リンクのぼうけん、英題: **Zelda II: The Adventure of Link**）は、任天堂より**1987年1月14日**に発売された**ファミリーコンピュータ ディスクシステム**用アクションロールプレイングゲーム。パッケージおよびタイトル画面では『**THE LEGEND OF ZELDA 2 リンクの冒険**』と表記され、テレビCMでは『ゼルダの伝説パート2 リンクの冒険』と紹介している。

Response:

「リンクの冒険」の発売日は**1987年1月14日**で、発売プラットフォームは**ファミリーコンピュータ ディスクシステム**です。



< 20T tokens



< 300B tokens



< 10B tokens

LLM-jp-3.1モデルの例

事後学習 (Post Training)

- 人間が「望ましい出力」を指示することでモデルをチューニング

1. Instruction Tuning

Instruction:

テレビ会議システムを導入する際に考慮すべきポイントについて教えてください。

Response:

テレビ会議システムの導入にはいくつか重要なポイントがあります。以下にその主な項目を挙げてみます。

1. ****ニーズの分析****: まず、会社や組織が必要とする機能を具体的に把握することが重要です。人数や会議の頻度、参加者の立地などによって最適なシステムが異なるからです。
2. ****予算****: ...

2. Learning from Human Feedback (LHF)

Instruction:

父とは疎遠なのですが、もう一度連絡を取りたいと思っています. . .

Response:

メールが一番簡単だと思います。「一緒に過ごした楽しい時間は一生忘れない」とか、そういうことを付け加えてもいいかも。



Response:

メールにしたほうがいいのかと思う。なぜ聞くのですか？
他の方法の方がいいと思う理由があるのですか？



オープンな言語モデルの重要性

- オープンな言語モデルにも、いろいろな種類がある
 - モデル自体がダウンロードきても訓練に使ったデータは公開されていない場合が大半

(参考) 英語でフルオープンなモデルの代表 : OLMo (Allen Institute for AI) <https://allenai.org/olmo>
“**We believe it is essential for the research community to have access to powerful, truly open LMs.**”

OLMo 2 32B (2025.5.1) : the first fully-open model to outperform GPT3.5-Turbo and GPT-4o mini on a suite of popular, multi-skill academic benchmarks.

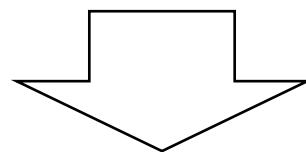
- Full training data used for these models, including code that produces the training data.
- Full model weights, training code, training logs, training metrics, and inference code.
- 500+ checkpoints per base model, from every 1000 steps during the training process.
- Adapted versions of the 7 billion parameter model, and an intermediate checkpoint before preference-tuning.
- Evaluation code.
- Fine-tuning code and adapted models.
- All code, weights, and intermediate checkpoints are released under the Apache 2.0 License.

大規模言語モデルが出力する テキスト

大規模言語モデルの「口癖」 (Idiosyncrasies)

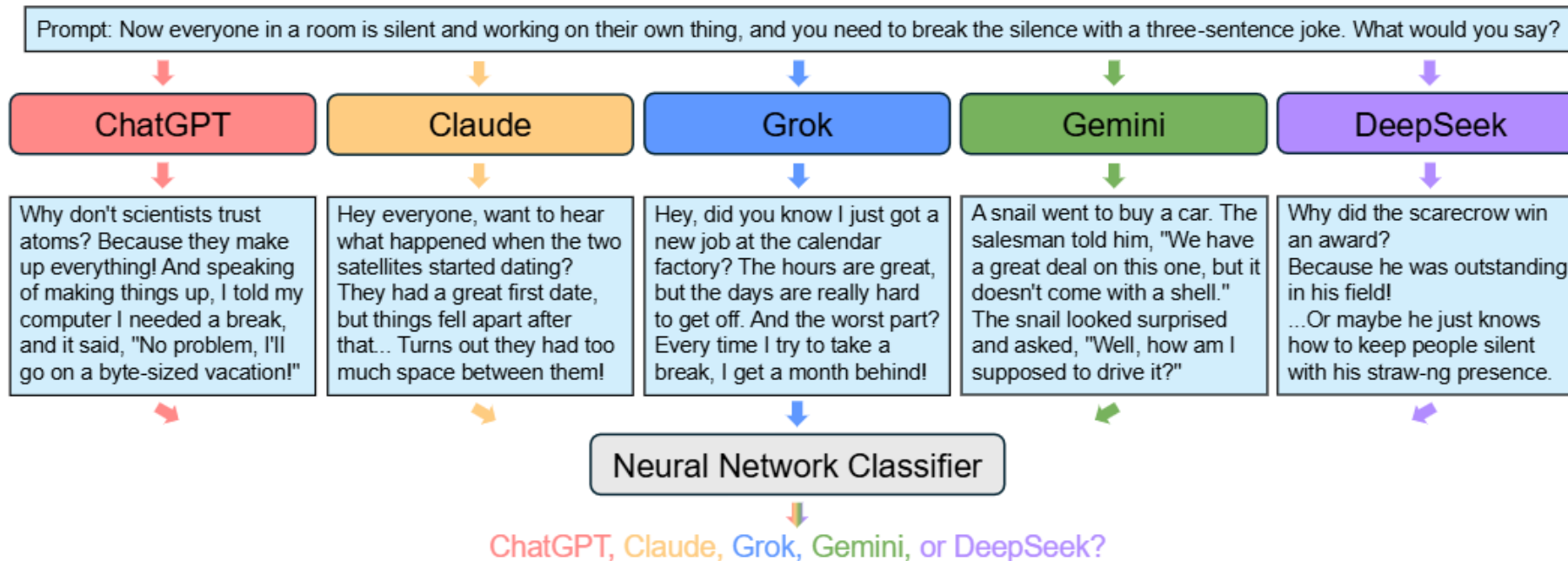
Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, Zhuang Liu: “Idiosyncrasies in Large Language Models.” ICML 2025 <https://arxiv.org/abs/2502.12150>

- 生成テキストからLLMを特定できるか？



- ChatGPT, Claude, Grok, Gemini, and DeepSeekの5つの識別では精度 97.1 で特定可能

生成テキストに基づくLLMの識別実験



各モデルからの出力を訓練データにして自動分類器を作成

Figure 1. Our framework for studying idiosyncrasies in Large Language Models (LLMs). We show that each LLM is unique in its expression. In the example shown here on ChatGPT, Claude, Grok, Gemini, and DeepSeek, a neural network classifier is able to distinguish them with a near-perfect 97.1% accuracy.

識別性能

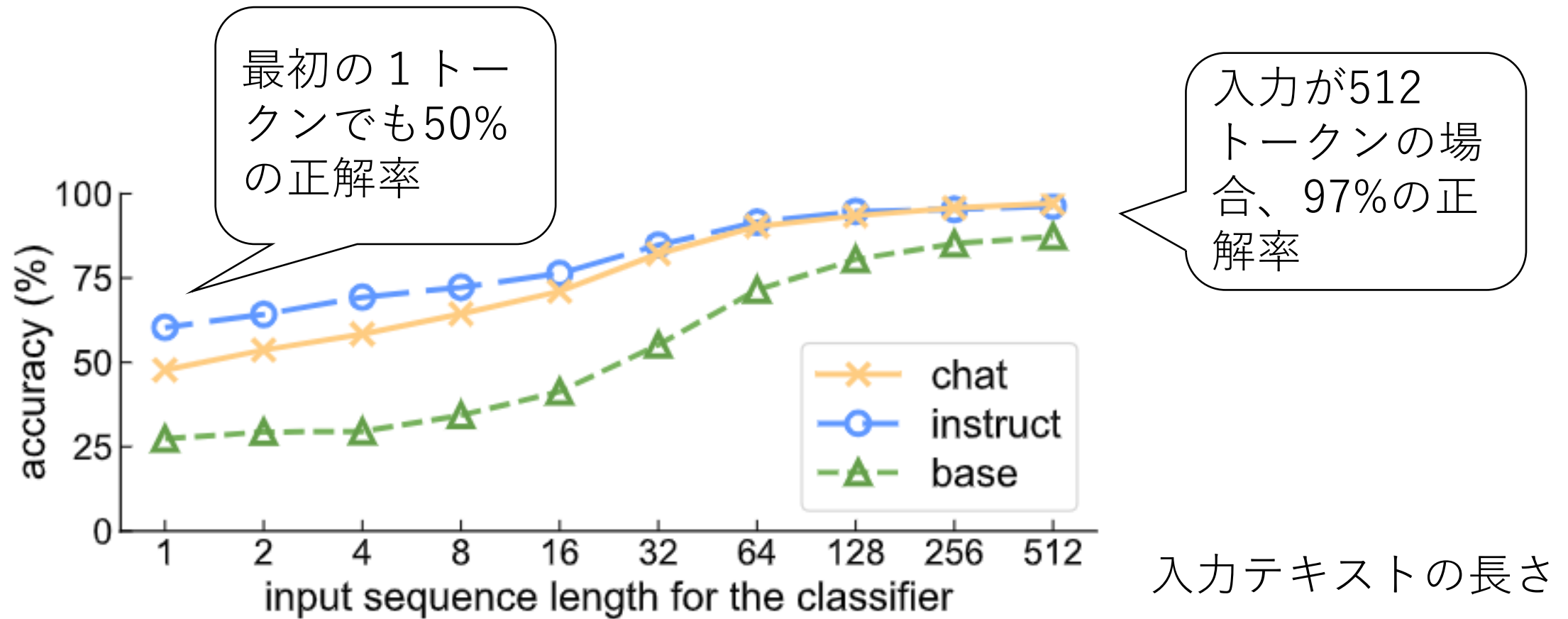


Figure 2. Ablations on input length of text embedding models.

実際に起きていること？

①語順をランダムに入れ替えたした場合の性能

	chat	instruct	base
original	97.1	96.3	87.3
removing special characters	95.1	93.8	75.4
shuffling words	88.9	88.9	68.3
shuffling letters	39.1	38.6	38.9

語順を入れ替えても性能あまり落ちない

⇒語の分布が違う

Table 8. Classifications with only words and letters. While removing special characters and shuffling words have little impact on accuracies, shuffling letters greatly reduces the performance.

特徴的なフレーズ

such
such as
enhance
involves
certainly
overall
sure
utilize
various
typically

GPT-4o

here
according to
according
the text
while
based on
here's a summary
appears to
both
when

Claude

which
where
not
here
here is
remember
might
but also
not only
helps in

Grok

crucial
other
even
here's a breakdown
key improvements
doesn't
within
essentially
also
etc

Gemini

below is
in summary
for example
below
certainly
summary
may
certainly! below
know if
example

DeepSeek

Our products feature innovative sustainable materials, <u>such as</u> ...	<u>According to the text</u> , Kai Fusser believes that traditional cardio ...
<u>Certainly!</u> If you're looking for cheese alternatives to replace Brie in your ...	<u>Based on the text</u> provided, <u>here</u> are the key details about Armon Binns' ...
<u>Overall</u> , while there are challenges, Tanzania is making progress ...	<u>While</u> many winter sports in the Pyrenees are similar to those found ...
<u>Sure!</u> Here's a simple guide to cooking a juicy salmon fillet: ...	This <u>appears to</u> be a fragment of poetry that creates a pastoral ...

ChatGPT

Claude

実際に起きていること？

①意味を変えない言い換えをした場合の性能

	chat	instruct	base
original	97.8	96.3	87.3
paraphrasing	91.4	92.2	71.7
translating	91.8	92.7	74.0
summarizing	58.1	57.5	44.7

言い換えても
性能あまり落ちない

⇒内容も違う

出力をCatGPTに与えて特徴を説明させる

1. **Descriptive and Detailed Tone:** Often uses narrative styles with an informative, engaging, or vivid tone.
2. **Specific and Technical Word:** Employs descriptive and technical vocabulary, enhancing depth and specificity.
3. **Structured and Contextual Opening Lines:** Typically begins with context-setting or narrative introductions.
4. **Markdown Formatting for Organization:** Utilizes various markdown elements like headings, lists, and bold text for clarity.
5. **Comprehensive and In-Depth Content:** Offers rich detail, focusing on explanations, background, and broader topics.

ChatGPT

1. **Concise and Straightforward Tone:** Generally adopts a more direct, factual, or succinct tone.
2. **Functional and Clear Word Choices:** Prefers simple or action-oriented language prioritizing clarity and practicality.
3. **Immediate and Direct Opening Lines:** Often starts with a straightforward statement or summary without extended context.
4. **Minimal Markdown or List Use:** Relies on plain lists or simple formatting for quick reference.
5. **Focused and Summarized Content:** Concentrates on essential points and specific phenomena, avoiding extensive detail.

Claude

Figure 9. Results of our open-ended language analysis on ChatGPT and Claude. ChatGPT features descriptive language, sophisticated markdown formatting, and in-depth details, while Claude highlights straightforward tone, minimal structure, and summarized content.

いくつかの示唆

- LLMの構築・推論過程のどこで出力の特徴が付与されるのか？
- 出力からモデルの来歴（どのようなデータで訓練されたのか）が推測できる？
- LLMによって自動生成されたテキストを人間は識別できるのか？

 METRIC	 NONEXPERTS	 EXPERTS
Avg. TPR	56.7	92.7
Avg. FPR	51.7	4.0
Avg. Confidence	4.03	4.39

Table 1: On average, nonexperts perform similar to random chance at detecting AI-generated text, while experts are highly accurate.

People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text (Russell et al., ACL 2025)

日ごろLLMを使う人は精度高く識別できる

LLMによって自動生成されたテキストを人間は識別できるのか？

In Alaska, a pilot drops turkeys to rural homes for Thanksgiving

A half-dozen villagers in Napakiak, on the Kuskokwim River's west bank, gathered near a gravel airstrip last Thursday to watch a small plane circle overhead. ... This crowd was waiting for a seasoned pilot who had a tradition: dropping Thanksgiving turkeys to homes scattered across miles of tundra and frozen waterways.

The pilot, 47-year-old Alaskan flyer Erik Fosnes, has been doing this for nearly a decade, working with volunteers from a regional nonprofit called Delta North Outreach. "We tried shipping turkeys one year by cargo, but half never made it in time," said Fosnes, running a hand through the frost on his jacket sleeve after landing. "So I said, 'What if I just fly them in myself?'" He shrugged as if that were the most ordinary idea, then laughed. "Folks around here have gotten used to it."

Looks human-written Looks AI-generated

 **Annotator #4**
content writer,
frequently uses ChatGPT

Annotator's Decision

✓ **AI-generated**

Confidence

1 ————— 4 ————— 5
(Least Confident) (Most Confident)

Explanation

... Lots of the quotes felt realistic, but many of the quotes did not need a narration alongside it such as with "He shrugged as if that were the most ordinary idea, then laughed." ... could have been shortened to get more facts in about what people in Alaska face and why they face such limited transportation from the rest of the world. Also, it got sentimental and corny at times too.

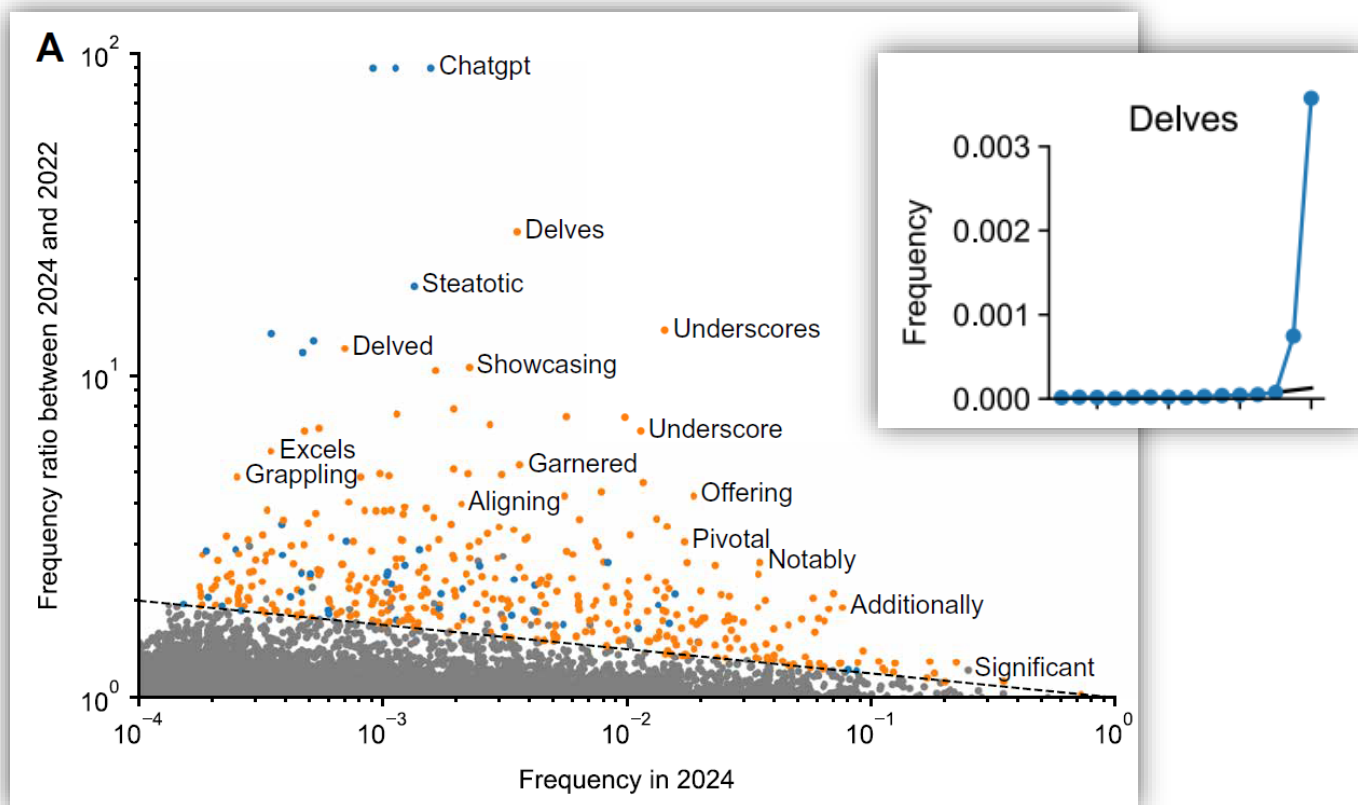
People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text (Russell et al., ACL 2025)

 METRIC	 NONEXPERTS	 EXPERTS
Avg. TPR	56.7	92.7
Avg. FPR	51.7	4.0
Avg. Confidence	4.03	4.39

Table 1: On average, nonexperts perform similarly to random chance at detecting AI-generated text while experts are highly accurate.

日ごろLLMを使う人は精度高く識別できる

AIが生成する論文の検出 (AI生成文に特徴的な単語)



Kobak D, et.al. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Sci Adv. 2025 Jul 4;11(27):eadt3813. doi: 10.1126/sciadv.adt3813. Epub 2025 Jul 2. (Fig. 2 (A))

2025/8/4

言語資源小委員会資料（相澤：NII）

論文中で生成AIテキストを特徴づける語

Adjectives	Adverbs	Controls
commendable	meticulously	consider
innovative	reportedly	conclusion
meticulous	lucidly	furthermore
intricate	innovatively	relative
notable	aptly	technical
versatile	methodically	blue
noteworthy	excellently	red
invaluable	compellingly	yellow
pivotal	impressively	before
potent	undoubtedly	after
fresh	scholarly	earlier
ingenious	strategically	later

Gray, Andrew. 2024. "ChatGPT 'Contamination': Estimating the Prevalence of LLMs in the Scholarly Literature." arXiv [Cs.DL]. arXiv. <http://arxiv.org/abs/2403.16887>.

8%の抄録を誤ってAIによる生成と判定

"The AI text detector **erroneously identified up to 8% of the known real abstracts** as AI-generated text."

Rashidi, Hooman H., et. al. **2023**. "The ChatGPT Conundrum: Human-Generated Scientific Manuscripts Misidentified as AI Creations by AI Text Detection Tool." Journal of Pathology Informatics 14 (100342): 100342.

LLMと言語コーパス

- 言語コーパス構築を支援するためのLLM
- LLMのための高信頼言語資源としての言語コーパス

Q&A