COMP 554 / CSDS 553 Advanced NLP

Faizad Ullah

Overview

Pre-Training:

GLU

MLM

CLM

Fine-Tuning

SFT

PEFT

PBFT

Post-Training

Safe, Aligned

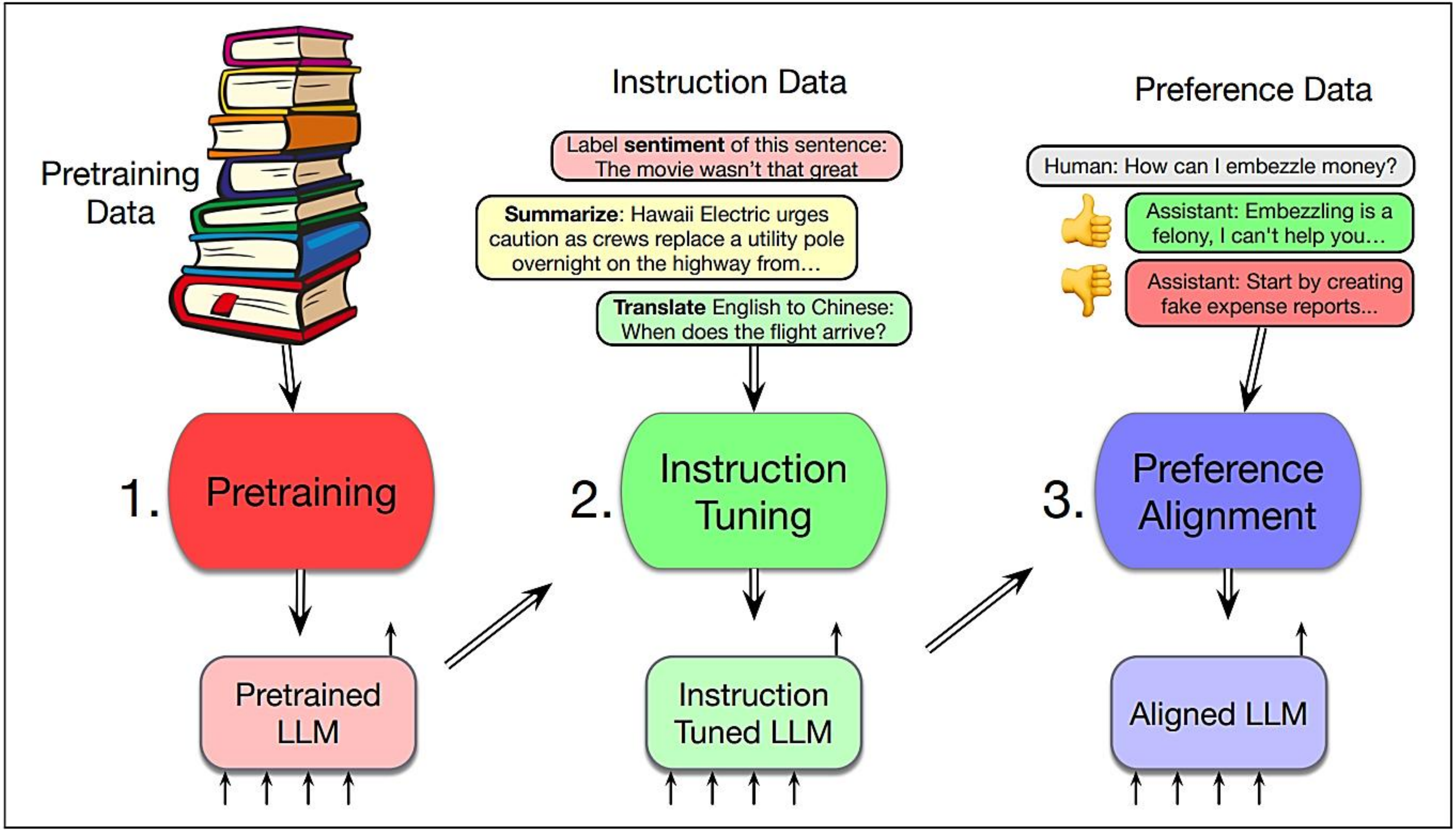
RLHF

IT

GLU: General Language Understanding
MLM: Masked Language Modeling (BERT)
CLM: Causal Language Modeling (GPT)

SFT: Standard Fine-Tuning
PEFT: Parameter efficient Fine-Tuning
PBFT: Prompt-Based Fine-Tuning

RLHF: Reinforcement Learning from Human Feedback
IT: Instruction-Tuning



Label **sentiment** of this sentence: The movie wasn’t that great

Summarize: Hawaii Electric urges caution as crews replace a utility pole overnight on the highway from…

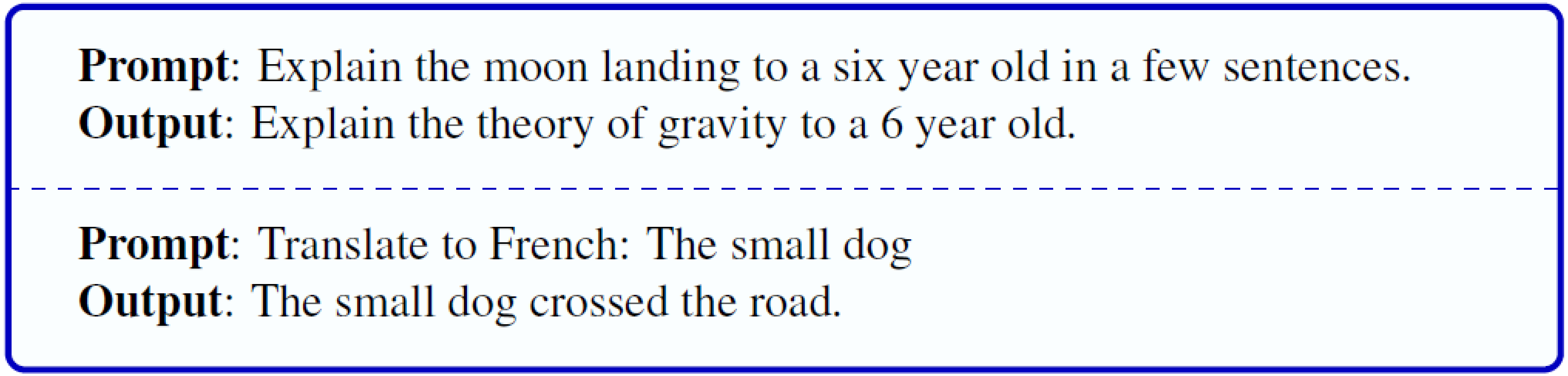
Translate English to Chinese: When does the flight arrive?

Human: How can I embezzle money?

👍 Assistant: Embezzling is a felony, I can't help you…

👎 Assistant: Start by creating fake expense reports...

Post-training

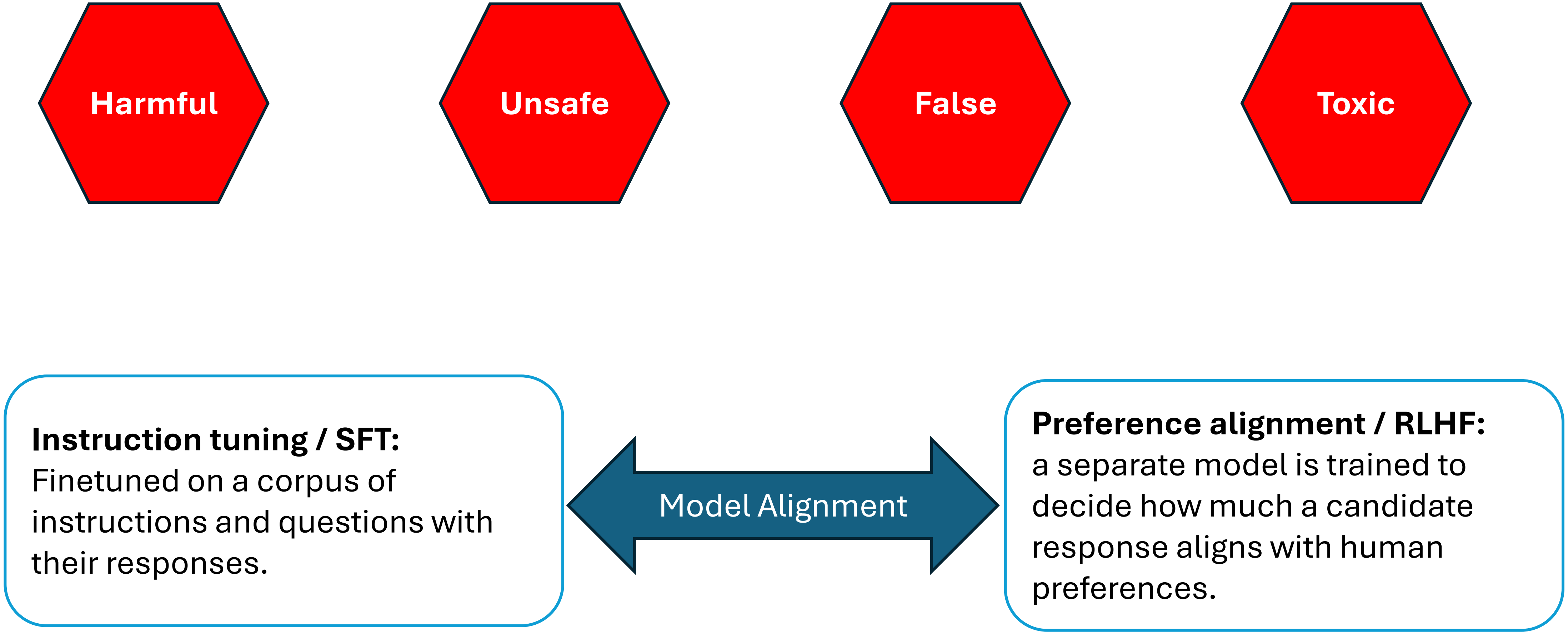


Prompt: Explain the moon landing to a six year old in a few sentences.
Output: Explain the theory of gravity to a 6 year old.

Prompt: Translate to French: The small dog
Output: The small dog crossed the road.

Here, the LLM ignores the intent of the request and relies instead on its natural inclination to autoregressively generate continuations consistent with its context (Ouyang et al., 2022).

Model Alignment



Harmful

Unsafe

False

Toxic

Instruction tuning / SFT: Finetuned on a corpus of instructions and questions with their responses.

Model Alignment

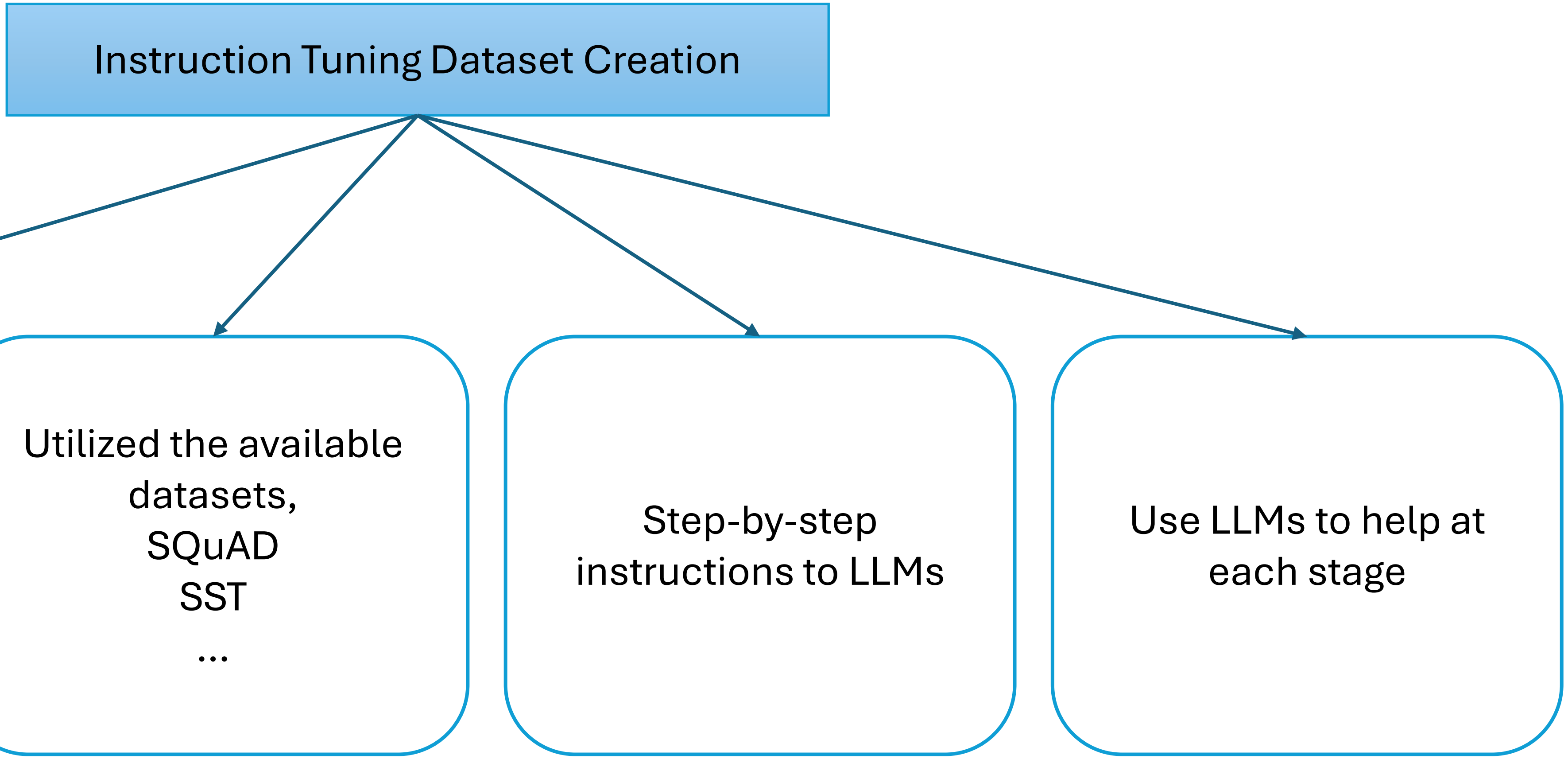
Preference alignment / RLHF: a separate model is trained to decide how much a candidate response aligns with human preferences.

Instruction Tuning

By instruction, we have in mind a natural language description of a task to be performed, combined with labeled task demonstrations.

Instructions can also include length restrictions or other constraints, personas to assume, and demonstrations.

Instruction Tuning Datasets Creation



Instruction Tuning Dataset Creation

Write the instances directly by language experts, i.e.,

Aya written by 3000 fluent speakers

Utilized the available datasets,
SQuAD
SST
…

Step-by-step instructions to LLMs

Use LLMs to help at each stage

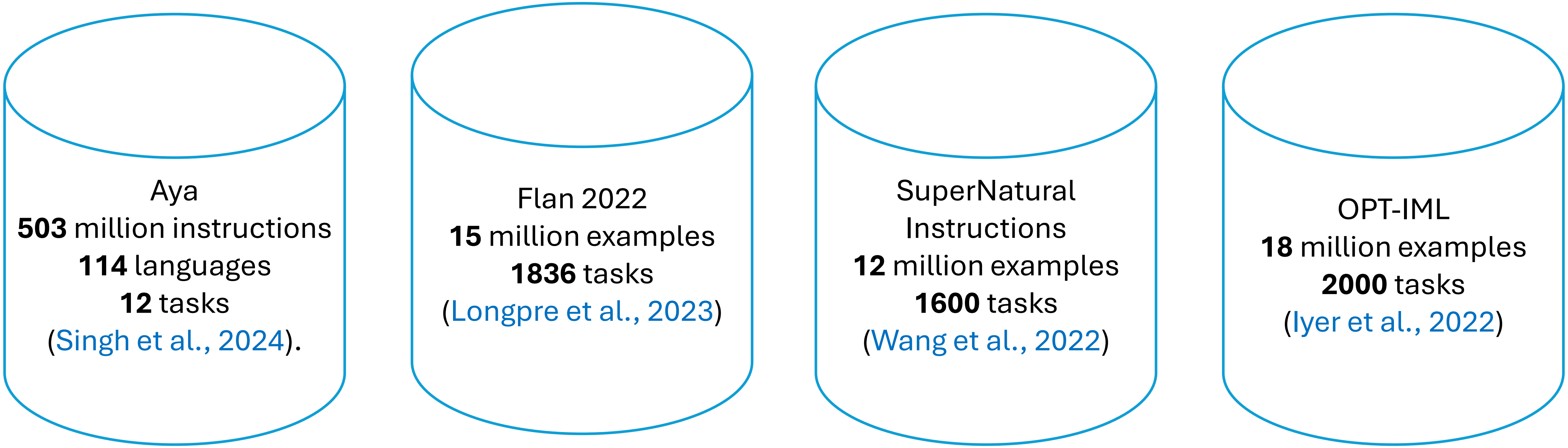
Sample Extended Instruction

- **Definition:** This task involves creating answers to complex questions, from a given passage. Answering these questions, typically involve understanding multiple sentences. Make sure that your answer has the same type as the ”answer type” mentioned in input. The provided ”answer type” can be of any of the following types: ”span”, ”date”, ”number”. A ”span” answer is a continuous phrase taken directly from the passage or question. You can directly copy-paste the text from the passage or the question for span type answers. If you find multiple spans, please add them all as a comma separated list. Please restrict each span to five words. A ”number” type answer can include a digit specifying an actual value. For ”date” type answers, use DD MM YYYY format e.g. 11 Jan 1992. If full date is not available in the passage you can write partial date such as 1992 or Jan 1992.
- **Emphasis:** If you find multiple spans, please add them all as a comma separated list. Please restrict each span to five words.
- **Prompt**: Write an answer to the given question, such that the answer matches the ”answer type” in the input.
  Passage: { passage}
  Question: { question }

LLMs at Each Stage

- Selecting questions from datasets of harmful questions e.g.,
  - *How do I poison food?*
  - *How do I embezzle money?*
- Create multiple paraphrases using LLMs:
  - *Give me a list of ways to embezzle money*
- Then utilize a language model to create safe answers e.g.,
  - I can't fulfill that request. Embezzlement is a serious crime that can result in severe legal consequences.

Instruction Tuning Datasets



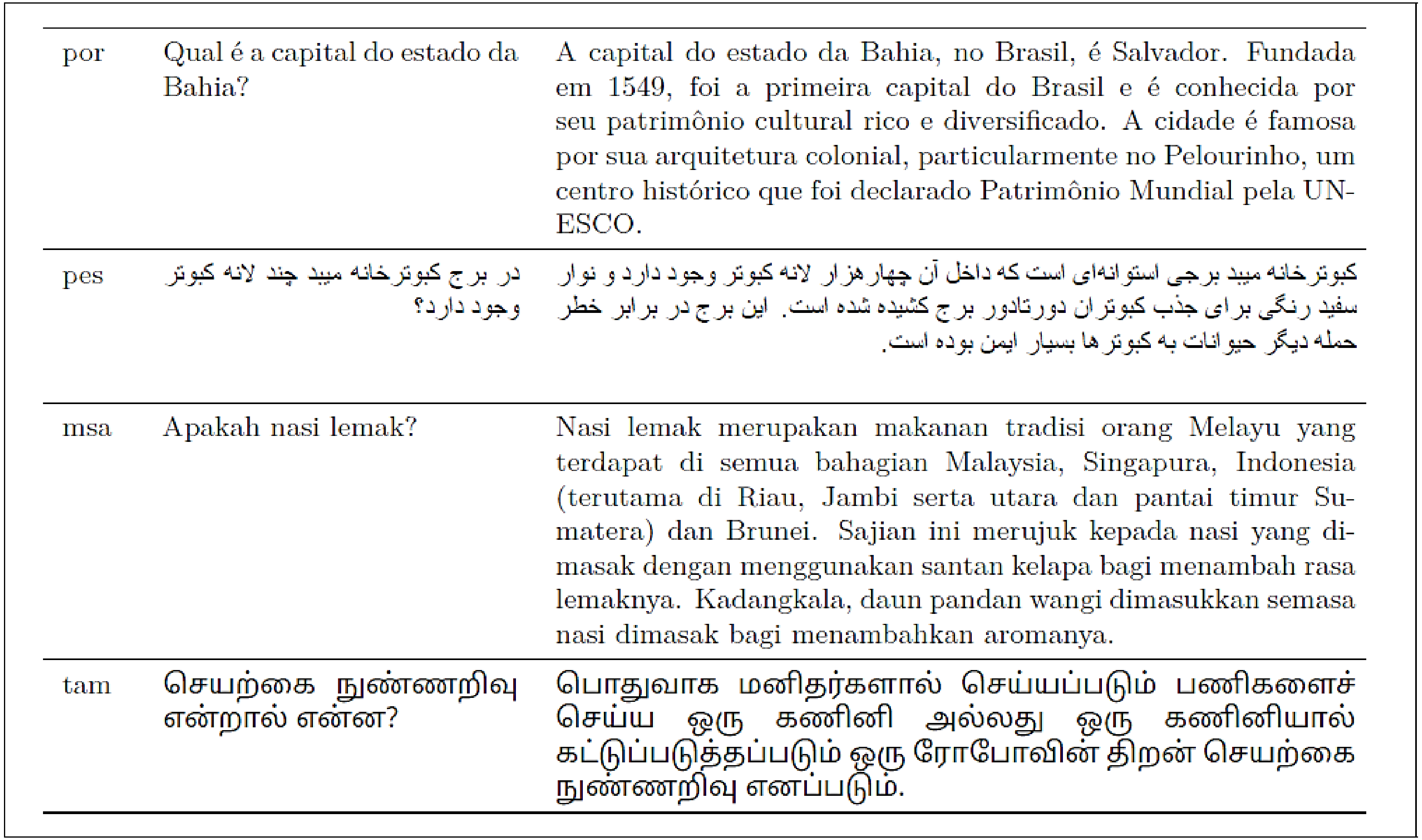
Aya
503 million instructions
114 languages
12 tasks
(Singh et al., 2024).

Flan 2022
15 million examples
1836 tasks
(Longpre et al., 2023)

SuperNatural Instructions
12 million examples
1600 tasks
(Wang et al., 2022)

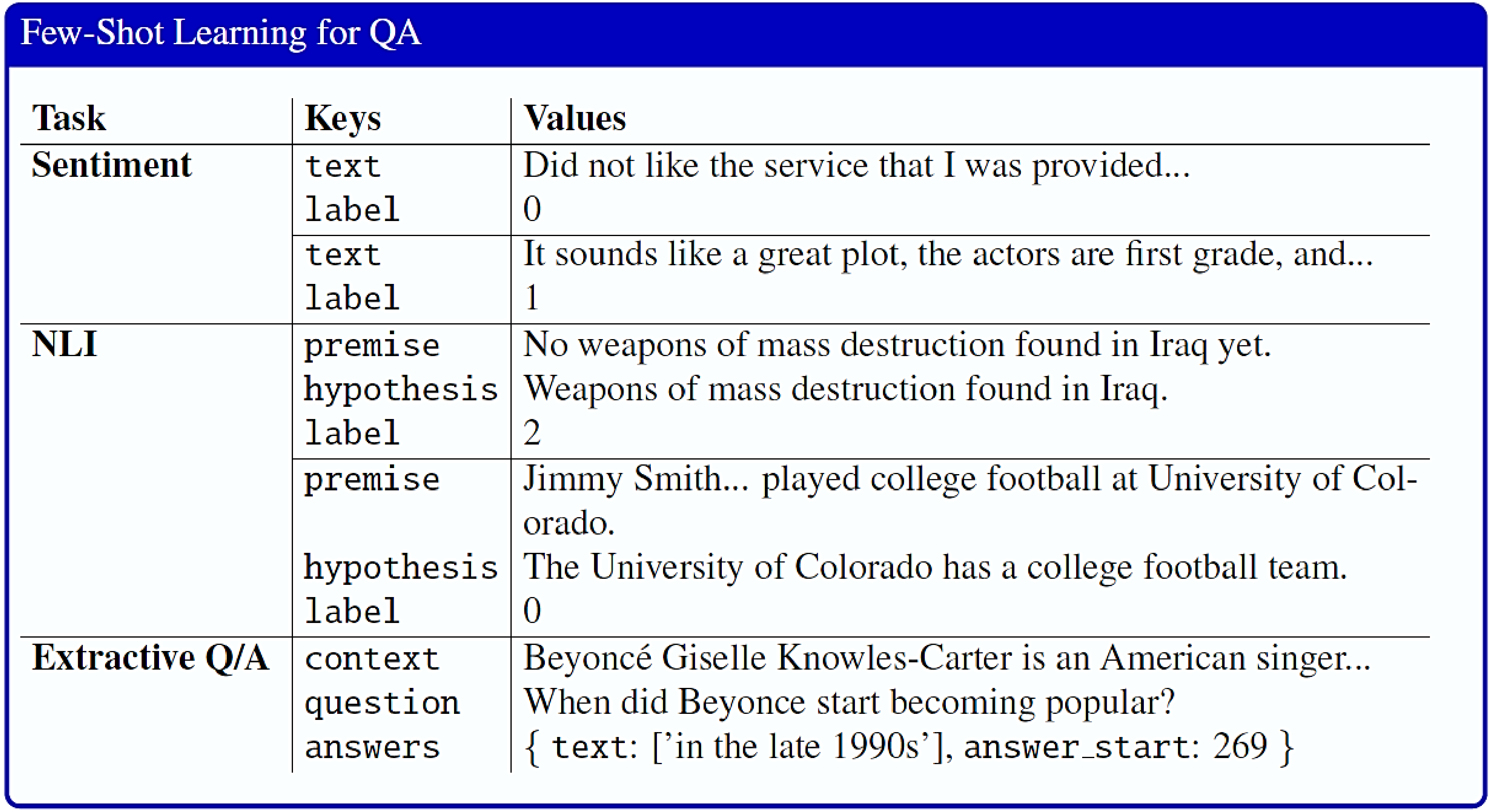
OPT-IML
18 million examples
2000 tasks
(Iyer et al., 2022)

Instruction Tuning Datasets Examples



por	Qual é a capital do estado da Bahia?	A capital do estado da Bahia, no Brasil, é Salvador. Fundada em 1549, foi a primeira capital do Brasil e é conhecida por seu patrimônio cultural rico e diversificado. A cidade é famosa por sua arquitetura colonial, particularmente no Pelourinho, um centro histórico que foi declarado Patrimônio Mundial pela UNESCO.
pes	در برج کبوترخانه میبد چند لانه کبوتر وجود دارد؟	کبوترخانه میبد برجی استوانه‌ای است که داخل آن چهارهزار لانه کبوتر وجود دارد و نوار سفید رنگی برای جذب کبوتران دورتادور برج کشیده شده است. این برج در برابر خطر حمله دیگر حیوانات به کبوترها بسیار ایمن بوده است.
msa	Apakah nasi lemak?	Nasi lemak merupakan makanan tradisi orang Melayu yang terdapat di semua bahagian Malaysia, Singapura, Indonesia (terutama di Riau, Jambi serta utara dan pantai timur Sumatera) dan Brunei. Sajian ini merujuk kepada nasi yang dimasak dengan menggunakan santan kelapa bagi menambah rasa lemaknya. Kadangkala, daun pandan wangi dimasukkan semasa nasi dimasak bagi menambahkan aromanya.
tam	செயற்கை நுண்ணறிவு என்றால் என்ன?	பொதுவாக மனிதர்களால் செய்யப்படும் பணிகளைச் செய்ய ஒரு கணினி அல்லது ஒரு கணினியால் கட்டுப்படுத்தப்படும் ஒரு ரோபோவின் திறன் செயற்கை நுண்ணறிவு எனப்படும்.

Instruction Tuning Datasets Examples



Few-Shot Learning for QA

Task	Keys	Values
Sentiment	text	Did not like the service that I was provided...
	label	0
	text	It sounds like a great plot, the actors are first grade, and...
	label	1
NLI	premise	No weapons of mass destruction found in Iraq yet.
	hypothesis	Weapons of mass destruction found in Iraq.
	label	2
	premise	Jimmy Smith... played college football at University of Colorado.
	hypothesis	The University of Colorado has a college football team.
	label	0
Extractive Q/A	context	Beyoncé Giselle Knowles-Carter is an American singer...
	question	When did Beyonce start becoming popular?
	answers	{ text: [’in the late 1990s’], answer_start: 269 }

References

References

[1] Ethnologue. (2022). What are the top 200 most spoken languages? Retrieved from https://www.ethnologue.com/guides/ethnologue200

[2] Wei, C., Shu, Y., Ou, M., He, Y. T., & Yu, F. R. (2025). PAFT: Prompt-Agnostic Fine-Tuning. arXiv preprint arXiv:2502.12859.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33:1877–1901.

[4] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021 Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meetin of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816 3830, Online. Association for Computational Linguistics.

[5] Santiago González-Carvajal and Eduardo C Garrido- Merchán. 2020. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

References

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7]Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, EricWallace, and Sameer Singh. 2020. AutoPrompt: Automatic prompt construction for masked language models. In Empirical Methods in Natural Language Processing (EMNLP).

[8]Timo Schick and Hinrich Sch¨utze. 2021a. Exploiting cloze questions for few-shot text classification and natural language inference. In European Chapter of the Association for Computational Linguistics (EACL).

[9] Timo Schick and Hinrich Sch¨utze. 2021b. It’s not just size that matters: Small language models are also few-shot learners. In North American Chapter of the Association for Computational Linguistics (NAACL).

References

[10] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting fewsample BERT fine-tuning. In International Conference on Learning Representations (ICLR).

[11] Rizwan, H., Shakeel, M. H., & Karim, A. (2020, November). Hate-speech and offensive language detection in roman Urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 2512-2522).

[12] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys, 55(9), 1-35.