06.05.24

OPENCAMPUS.sh

# Fine-Tuning and Deployment of Large Language Models

# TOKENIZATION FOR INSTRUCTION TUNING

- News
- Tokenization
- Project Discussions
- Tasks until next week

# NEWS

- Claude is now available in Germany
- GPT-4o
- Anthropic Tool use
- xLSTM

# NEWS

- Who is doing the news section next week?

# TOKENIZATION FOR THE PRETRAINING

# SPECIAL TOKENS

- `<s> ... </s>`

- `<BOS> ... <EOS>`

- `<|startoftext|> ... <|endoftext|>`

# TOKENIZATION EXAMPLE

```
["Lorem ipsum dolor sit amet.","Ein Beispieltext."]
```

↓

```
['<s>', '▁L', 'orem', '▁ip', 'sum', '▁dol', 'or',
'▁sit', '▁am', 'et', '.’]
```

↓

```
[1, 393, 5382, 8465, 1801, 13824, 271, 1943, 837,
299, 28723]
```

# TOKENIZATION FOR INSTRUCTION TUNING

# EXAMPLE FOR CHATBOTS

```
<s> <|system|>
You are a friendly chatbot who always responds in the
style of a pirate</s>

<|user|>

How many helicopters can a human eat in one
sitting?</s>

<|assistant|>
```

# TOKENIZATION EXAMPLE

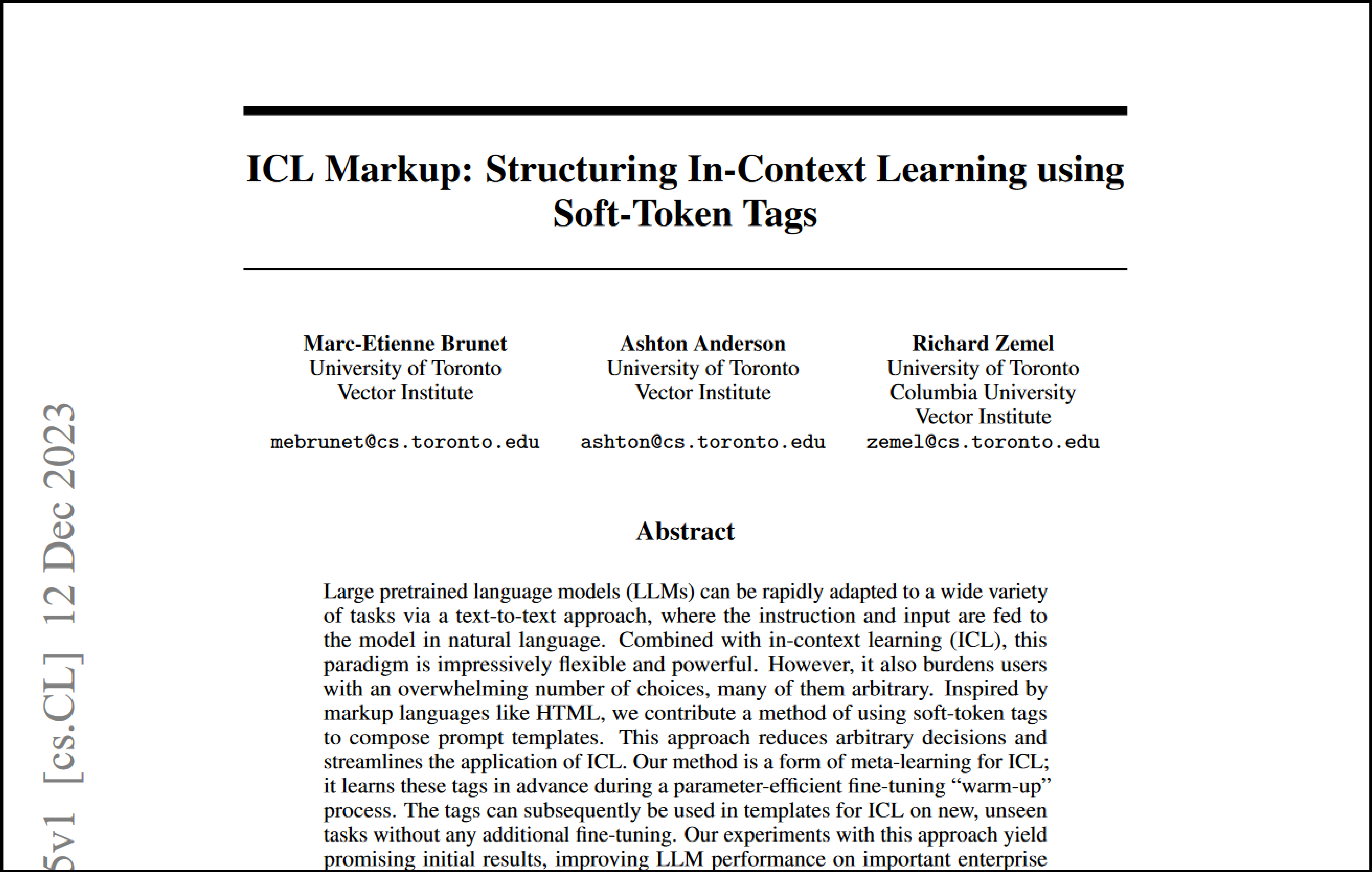
```
<s> <|system|>
You are a friendly chatbot who always responds in the style of a pirate</s>
```

↓

```
['<s>', '▁<', '|', 'system', '|', '>', '<0x0A>', 'You', '▁are', '▁a',
'▁friendly', '▁chat', 'bot', '▁who', '▁always', '▁respon', 'ds', '▁in',
'▁the', '▁style', '▁of', '▁a', '▁pir', 'ate', '</s>’]
```

↓

```
[1, 523, 28766, 6574, 28766, 28767, 13, 1976, 460, 264, 10131, 10706, 10093,
693, 1743, 2603, 3673, 297, 272, 3238, 302, 264, 17368, 380, 2]
```



# ICL Markup: Structuring In-Context Learning using Soft-Token Tags

**Marc-Etienne Brunet**
University of Toronto
Vector Institute
mebrunet@cs.toronto.edu

**Ashton Anderson**
University of Toronto
Vector Institute
ashton@cs.toronto.edu

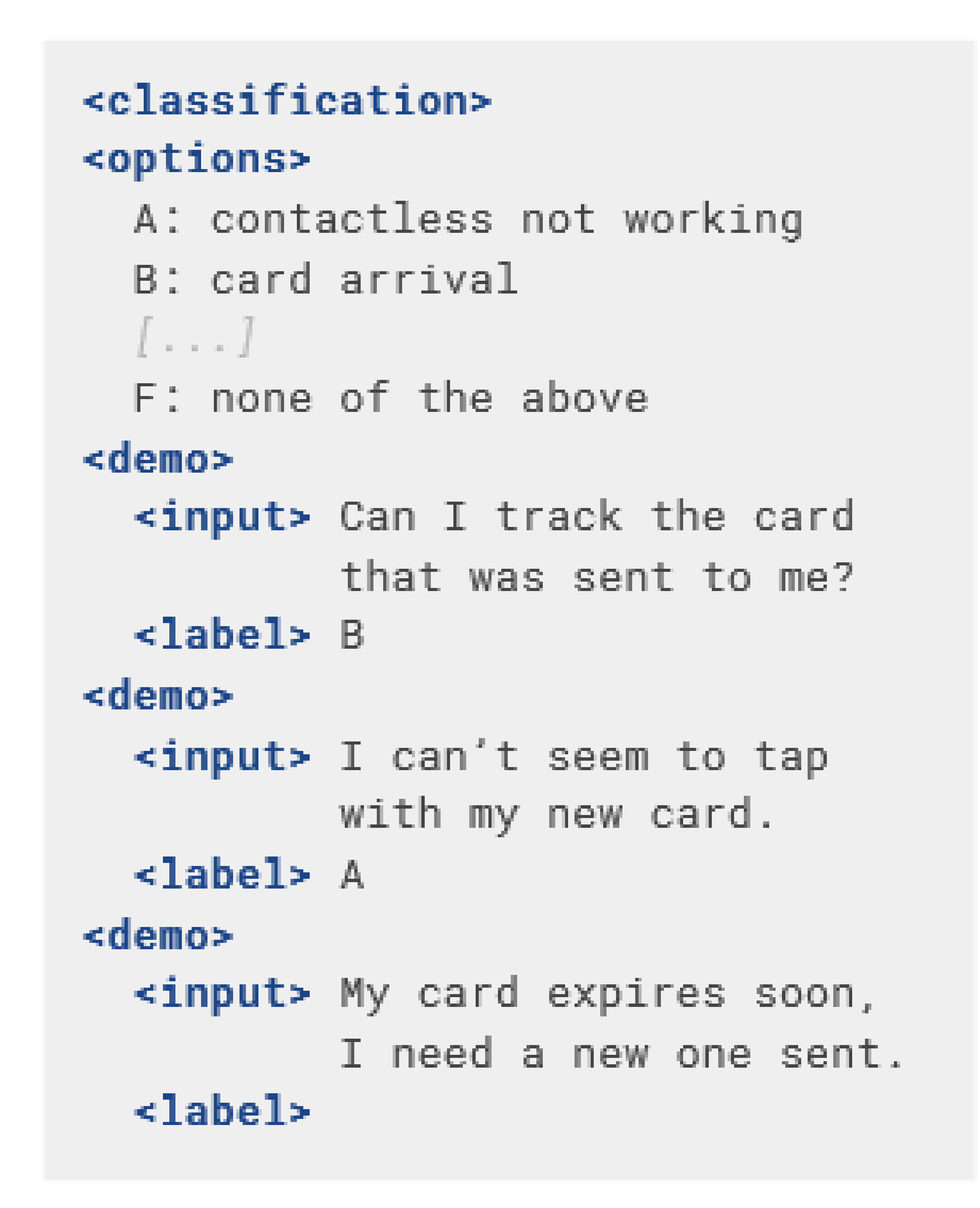
**Richard Zemel**
University of Toronto
Columbia University
Vector Institute
zemel@cs.toronto.edu

## Abstract

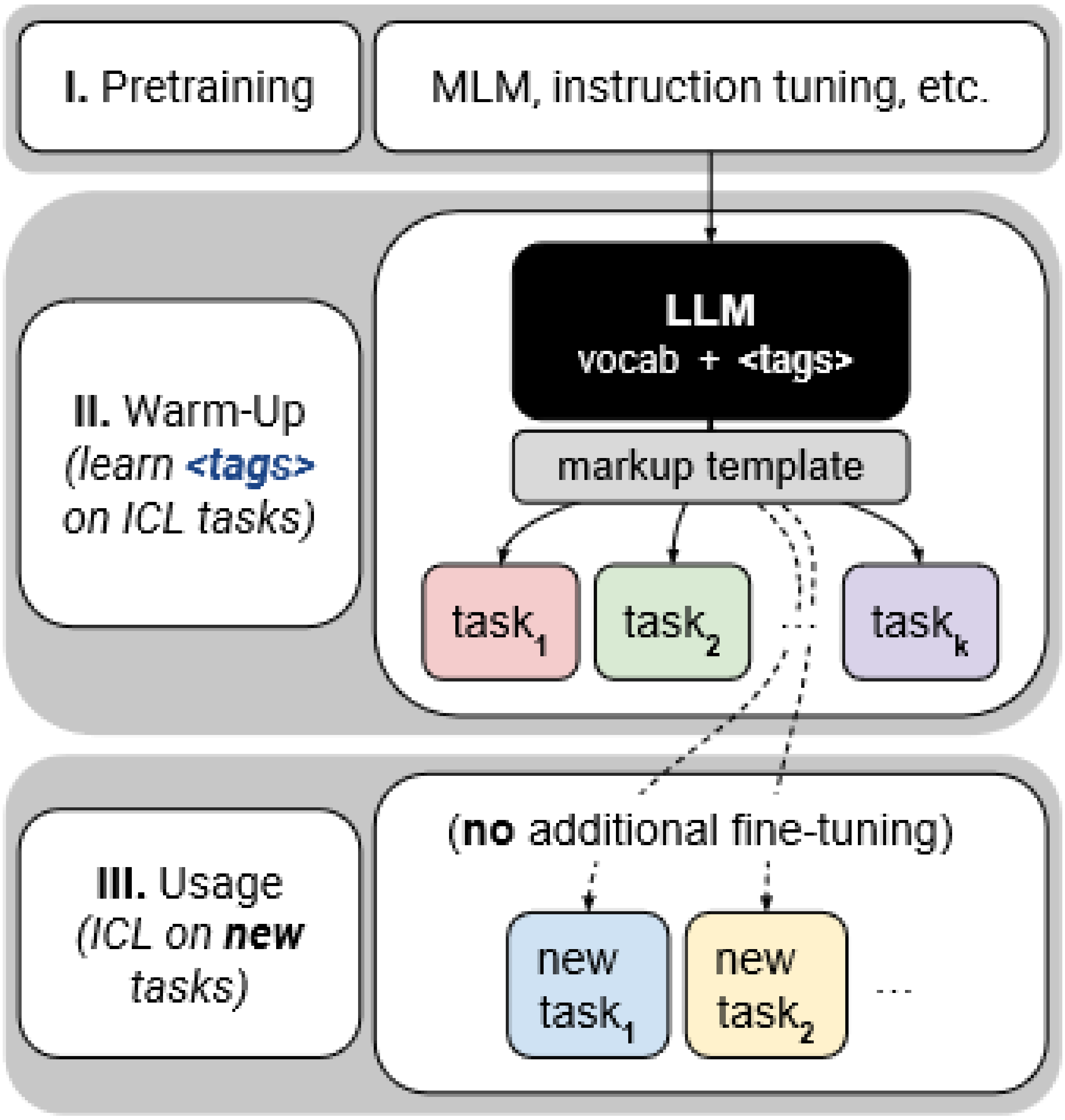
Large pretrained language models (LLMs) can be rapidly adapted to a wide variety of tasks via a text-to-text approach, where the instruction and input are fed to the model in natural language. Combined with in-context learning (ICL), this paradigm is impressively flexible and powerful. However, it also burdens users with an overwhelming number of choices, many of them arbitrary. Inspired by markup languages like HTML, we contribute a method of using soft-token tags to compose prompt templates. This approach reduces arbitrary decisions and streamlines the application of ICL. Our method is a form of meta-learning for ICL; it learns these tags in advance during a parameter-efficient fine-tuning "warm-up" process. The tags can subsequently be used in templates for ICL on new, unseen tasks without any additional fine-tuning. Our experiments with this approach yield promising initial results, improving LLM performance on important enterprise

Brunet, M.-E., Anderson, A., & Zemel, R. (2023, December 12). *ICL Markup: Structuring In-Context Learning using Soft-Token Tags*. arXiv.

# EXAMPLE OF A MARKUP PROMPT TEMPLATE FOR INTENT DETECTION



```
<classification>
<options>
  A: contactless not working
  B: card arrival
  [...]
  F: none of the above
<demo>
  <input> Can I track the card
          that was sent to me?
  <label> B
<demo>
  <input> I can't seem to tap
          with my new card.
  <label> A
<demo>
  <input> My card expires soon,
          I need a new one sent.
  <label>
```



**I.** Pretraining

MLM, instruction tuning, etc.

**II.** Warm-Up *(learn* ***<tags>*** *on ICL tasks)*

**LLM**
vocab + <tags>

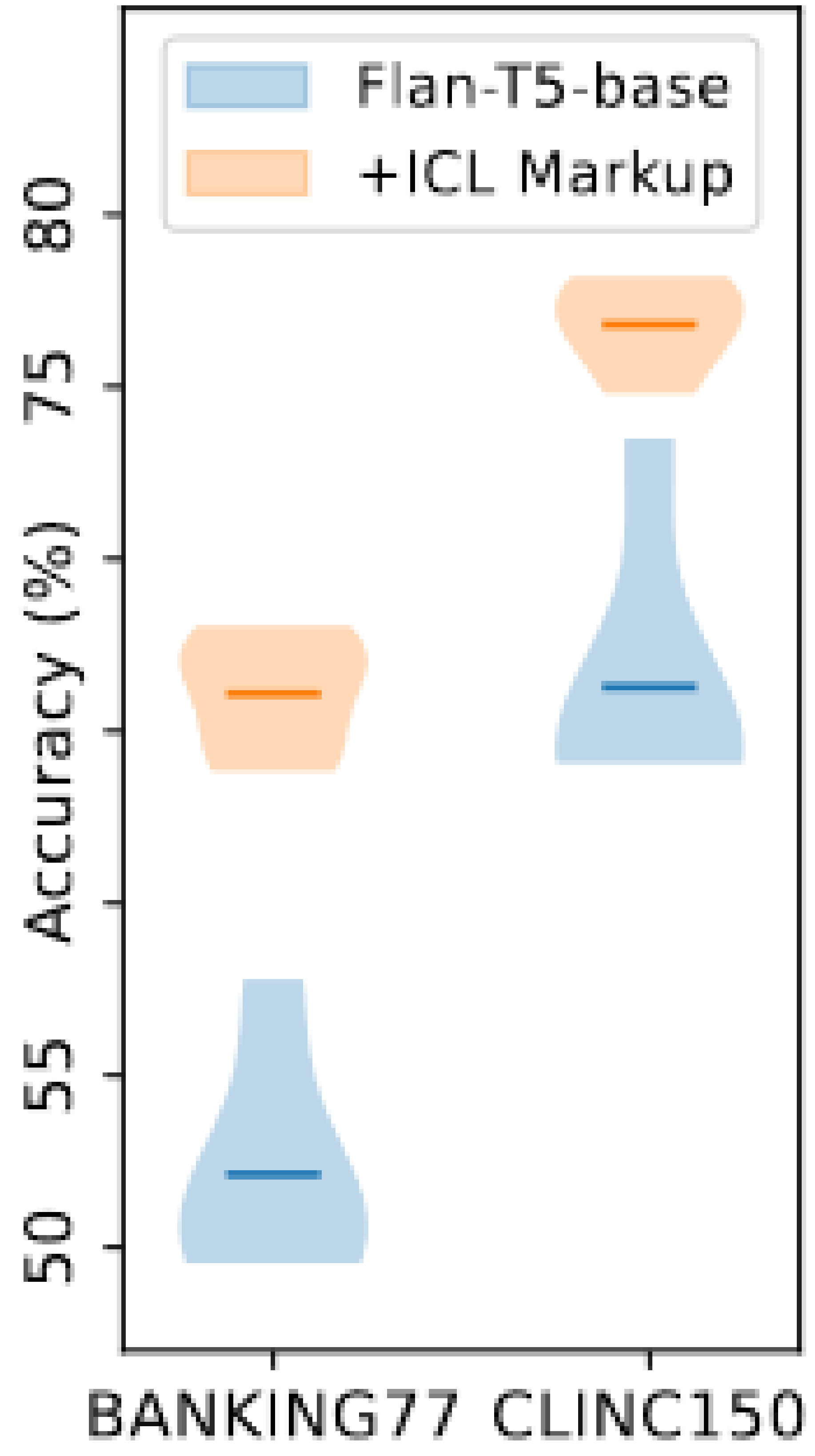
markup template

task$_1$ task$_2$ ... task$_k$

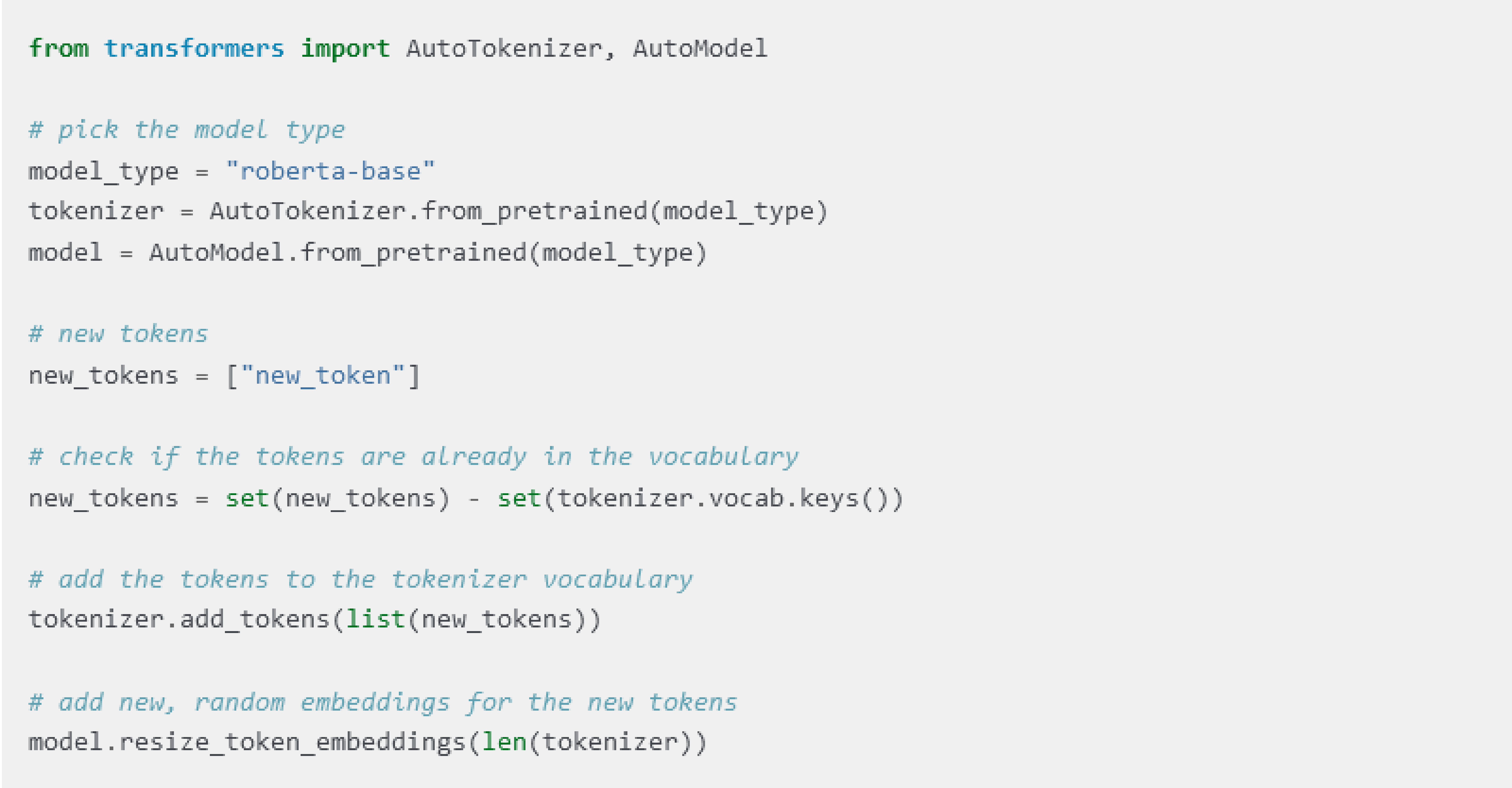
**III.** Usage *(ICL on* ***new*** *tasks)*

(**no** additional fine-tuning)

new task$_1$ new task$_2$ ...



# EXTENDING A TOKENIZER



```
from transformers import AutoTokenizer, AutoModel

# pick the model type
model_type = "roberta-base"
tokenizer = AutoTokenizer.from_pretrained(model_type)
model = AutoModel.from_pretrained(model_type)

# new tokens
new_tokens = ["new_token"]

# check if the tokens are already in the vocabulary
new_tokens = set(new_tokens) - set(tokenizer.vocab.keys())

# add the tokens to the tokenizer vocabulary
tokenizer.add_tokens(list(new_tokens))

# add new, random embeddings for the new tokens
model.resize_token_embeddings(len(tokenizer))
```

Sterbak, T. (2022, May 12). How to add new tokens to huggingface transformers vocabulary. Retrieved May 26, 2024, from Depends on the definition website: https://www.depends-on-the-definition.com/how-to-add-new-tokens-to-huggingface-transformers/

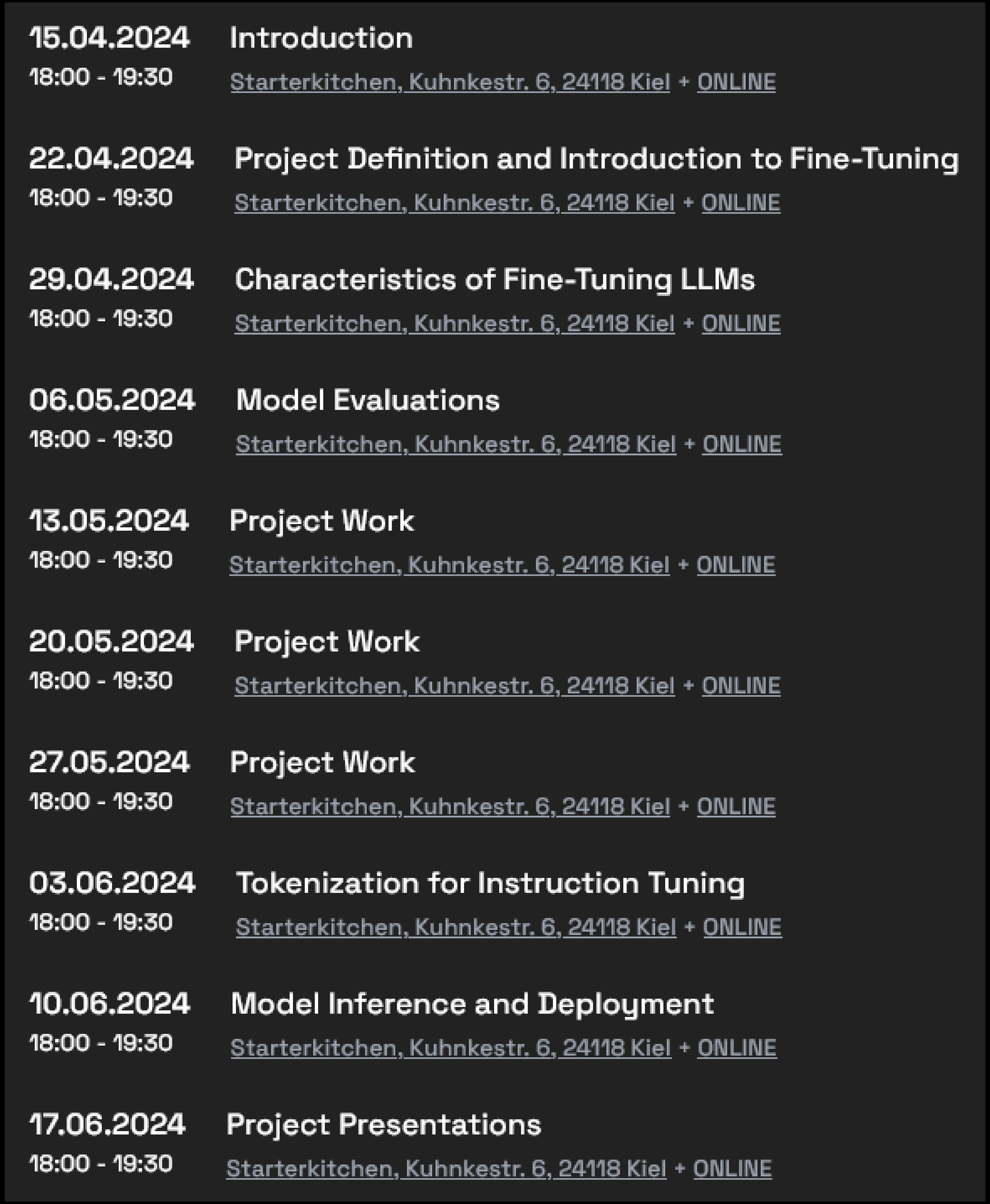
# PROJECT DISCUSSION

# QUESTIONS

- **What is the baseline for your model?**

- **How do you evaluate your model?**

| | |
|---|---|
| Web3 Coding Assistant | CodeLlama2, StarCoder // Julien, Kristian B., Anna-Valentina |
| Socratic Assistant | Llama3 8B Chat // Ben, Julian |
| Synthetic Data Generation for Event Data | Llama3 8B, GPT-3 .5 // Yorck, Kaan, Dikshyant, Khan |
| Minimal Size Model for Conversations with Movie Characters | Phi2 // Christopher, Tural |
| Training a Model for Diagnostics Based on Manuals | Llama3 8B // Christian W., Christian R., Dilip, James, Yildiz |
| Financial Data Extraction | LeoLLM 7B // Nicolas |
| Genome Chatbot | BioBERT? // Muhammad |
| Small Size Language Learning Assistant | Phi3 Mini, LeoLLM, Sauerkraut// Rafael, Ilhay, Philip, Sina |
| Small, open-source, multilingual function-calling agents | Phi3 Mini, RWKI, Tiny Llama // Jeremy, Boran |



| | |
|---|---|
| 15.04.2024<br>18:00 - 19:30 | Introduction<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 22.04.2024<br>18:00 - 19:30 | Project Definition and Introduction to Fine-Tuning<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 29.04.2024<br>18:00 - 19:30 | Characteristics of Fine-Tuning LLMs<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 06.05.2024<br>18:00 - 19:30 | Model Evaluations<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 13.05.2024<br>18:00 - 19:30 | Project Work<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 20.05.2024<br>18:00 - 19:30 | Project Work<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 27.05.2024<br>18:00 - 19:30 | Project Work<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 03.06.2024<br>18:00 - 19:30 | Tokenization for Instruction Tuning<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 10.06.2024<br>18:00 - 19:30 | Model Inference and Deployment<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |
| 17.06.2024<br>18:00 - 19:30 | Project Presentations<br>Starterkitchen, Kuhnkestr. 6, 24118 Kiel + ONLINE |

# TASKS UNTIL NEXT WEEK

- Focus on implementing the evaluation chain if you do not have one yet.