COLX585

What's behind the MASK?

— A Chinese Idiom Cloze Test
Using Fine-tuning and Prompt Engineering

Group Zootopia
Jessie, Gordon, Sherry, Dave, Grace

MOTIVATION

- Semantic analysis can be challenging for Chinese language corpus:

    - Ambiguity

    - Polysemy

    - Contextual information

    - Near-synonyms

    - …

- Most transformer-based models are trained on English corpus.

- We are linguistic students, let's try bilingual tasks.

Chinese idioms

- Also known as "Chengyu". Succinctly paraphrasing or summarizing the original text.
- A type of Chinese expression, most of which consist of four characters.
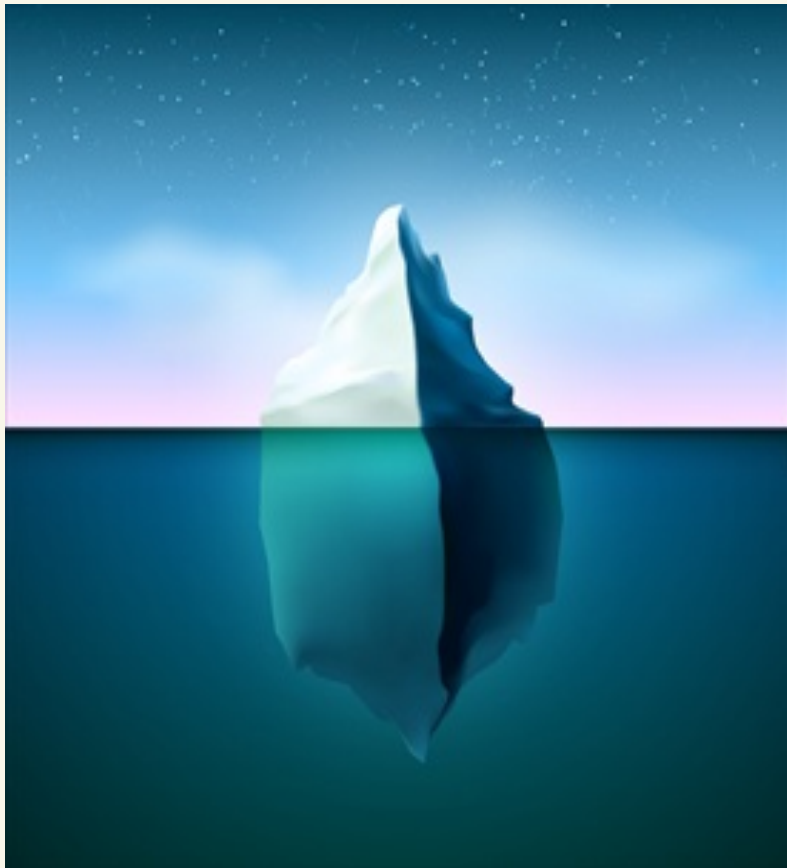
English:

one corner of an ice mountain

Chinese:

冰　　山　　一　　角

ice　　mountain　　one　　corner

Meaning：

Only a small portion of a larger issue or problem is visible, while the majority of the issue or problem remains hidden or unknown.

Chinese idioms

- Also known as "Chengyu". Succinctly paraphrasing or summarizing the original text.
- A type of Chinese expression, most of which consist of four characters.
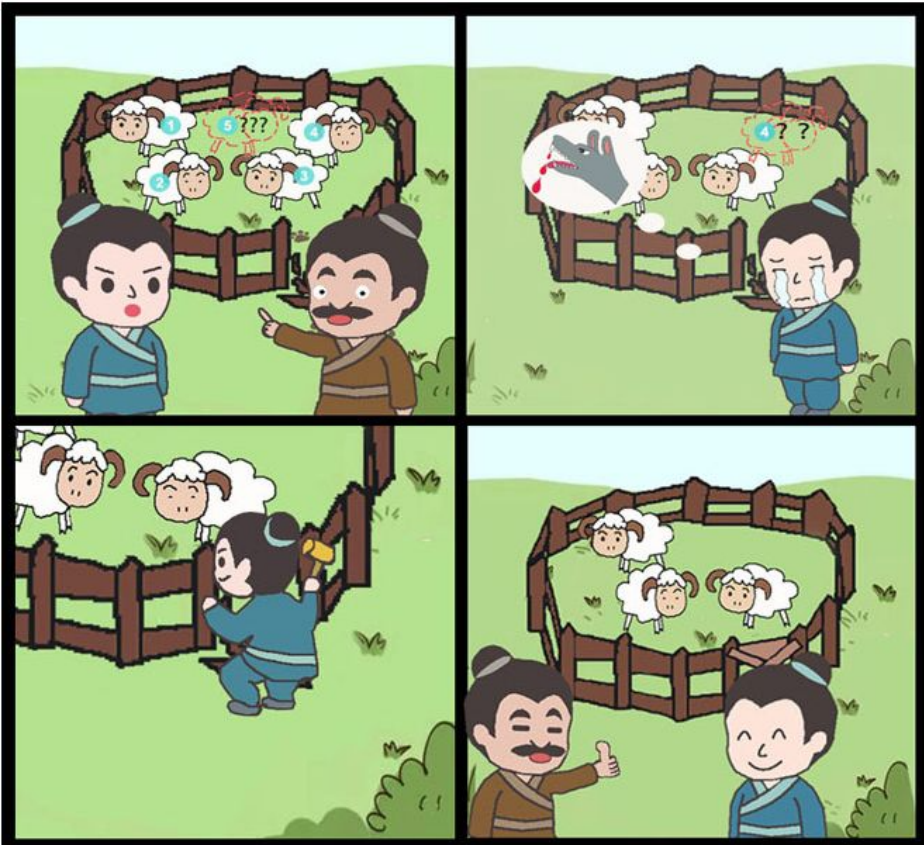
English:

To mend the fence after sheep are lost

Chinese:

亡　　　羊　　　补　　　牢

escape　　sheep　　mend　　fence

Meaning：

Never be late to make up

Note: An example of metaphor in idiom.

Computational linguistics, get the job done!

Dataset

- ChID, a Chinese cloze test dataset.

- Contains 581K passages and 729K blanks, and covers multiple domains.

- We use:

  train set: 15K

  dev set: 2K

  test set: 2K

Passage & Blanks		可是有一个时期大家#idiom-0#，不大敢露面，只有她一个人倚在阳台上看排队的兵走过。 However, there was a period when everyone *#idiom-0#* and was scared to show up. Only she leaned on the balcony and watched the soldiers passing by.
#idiom-0# options	**Correct**	深居简出　Be unwilling to contact people
	Similar	销声匿迹　To disappear from the scene 离群索居　To stay away from the crowd and live alone 安分守己　To know one's place
	Random	一帆风顺　To proceed smoothly without a hitch 文不对题　Be irrelevant to the subject 万里长征　A long and arduous undertaking

{"groundTruth": ["不寒而栗"],

 "candidates": [["瞠目结舌", "不寒而栗", "骇人听闻", "大功告成", "望而生畏", "时不我待"]],

"content": 个人信息一旦被公之于众，生活随之将被严重骚扰，这对当事人造成的心理压力可想而知，而因隐私泄露所引发的潜在危险也令人#idiom#。",

 "realCount": 1}

Computational linguistics, get the job done!

Previous work

- **Chujie et al. (2020)** evaluated three state-of-art cloze test models on ChID corpus, and observed that existing model performance is still much worse than human performance.

- **Ying et al. (2023)** experiment results on ChID shows that their model out performs state-of-art models. The accuracy on dataset Test reaches 95.7%, and the accuracy on dataset Sim, which has more synonyms or near-synonyms idioms, is significantly improved to 97.9%.
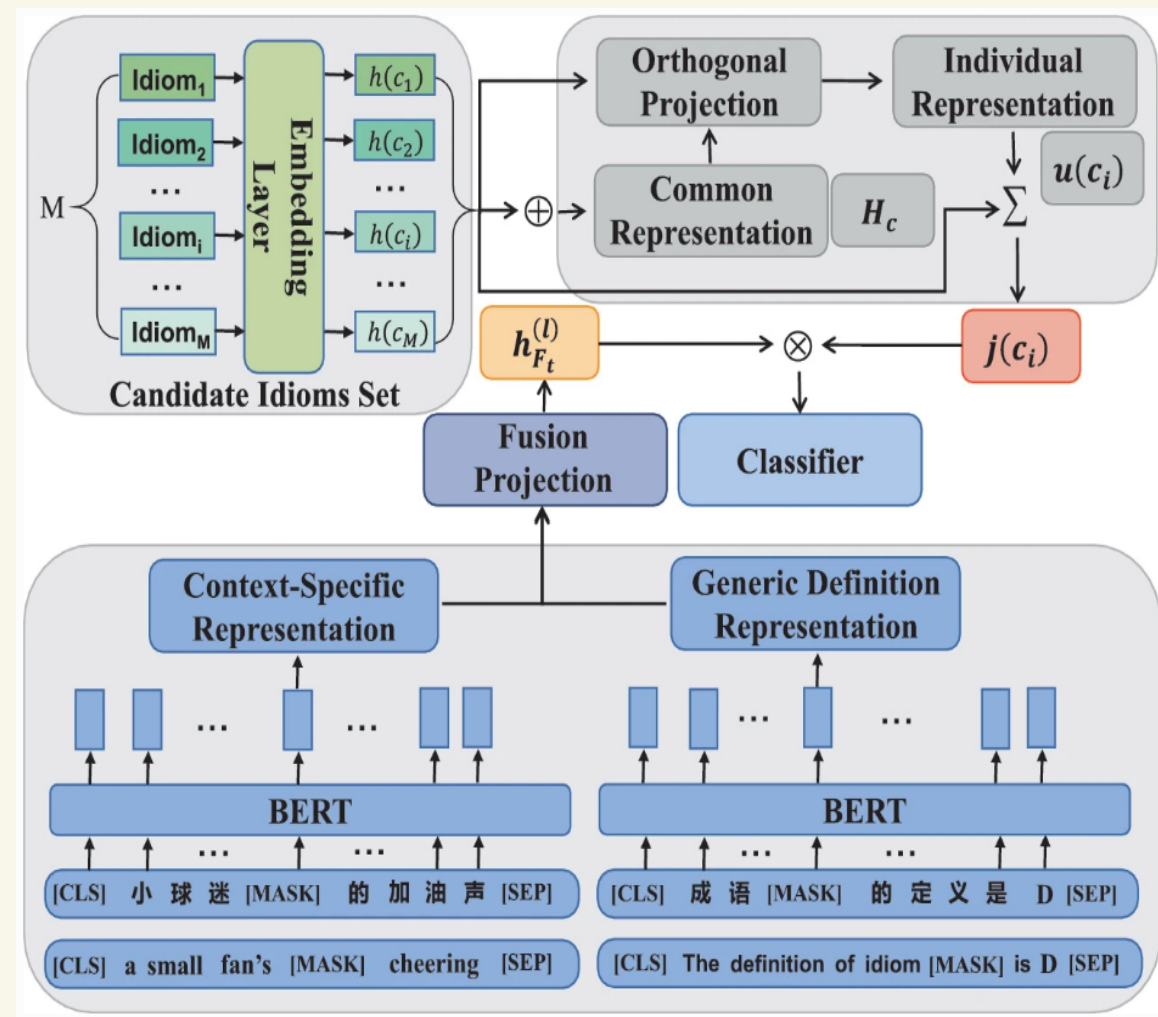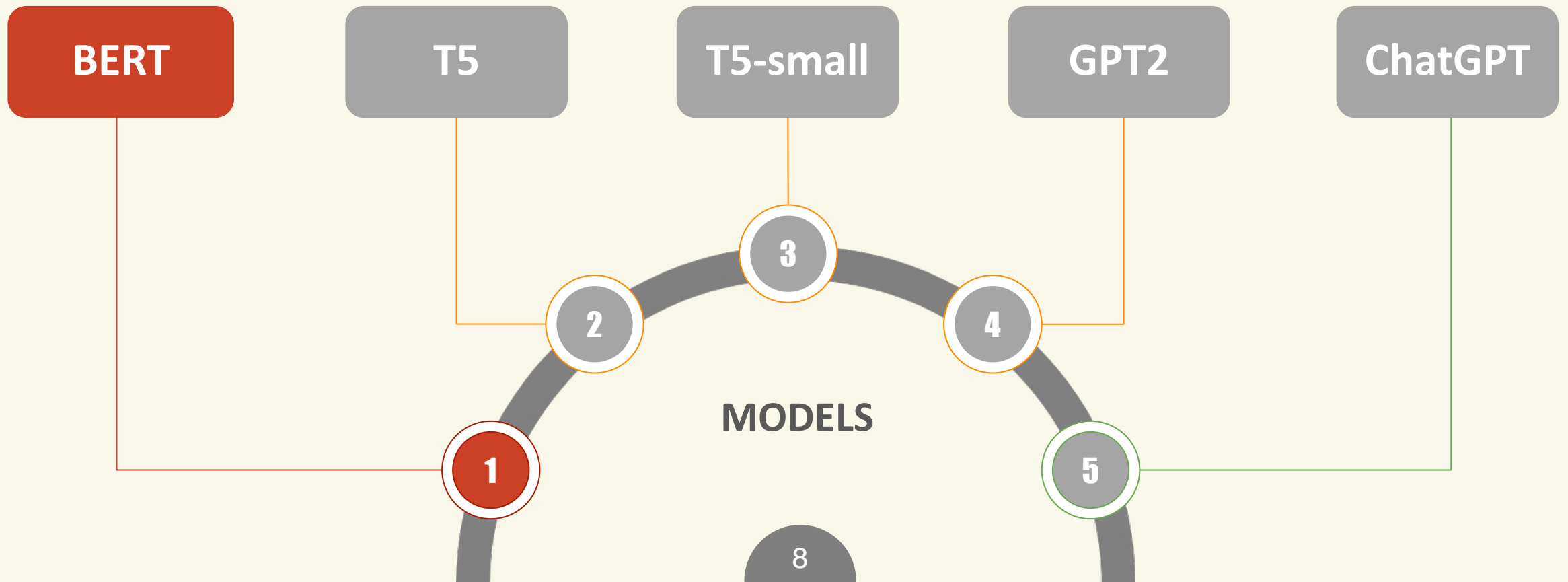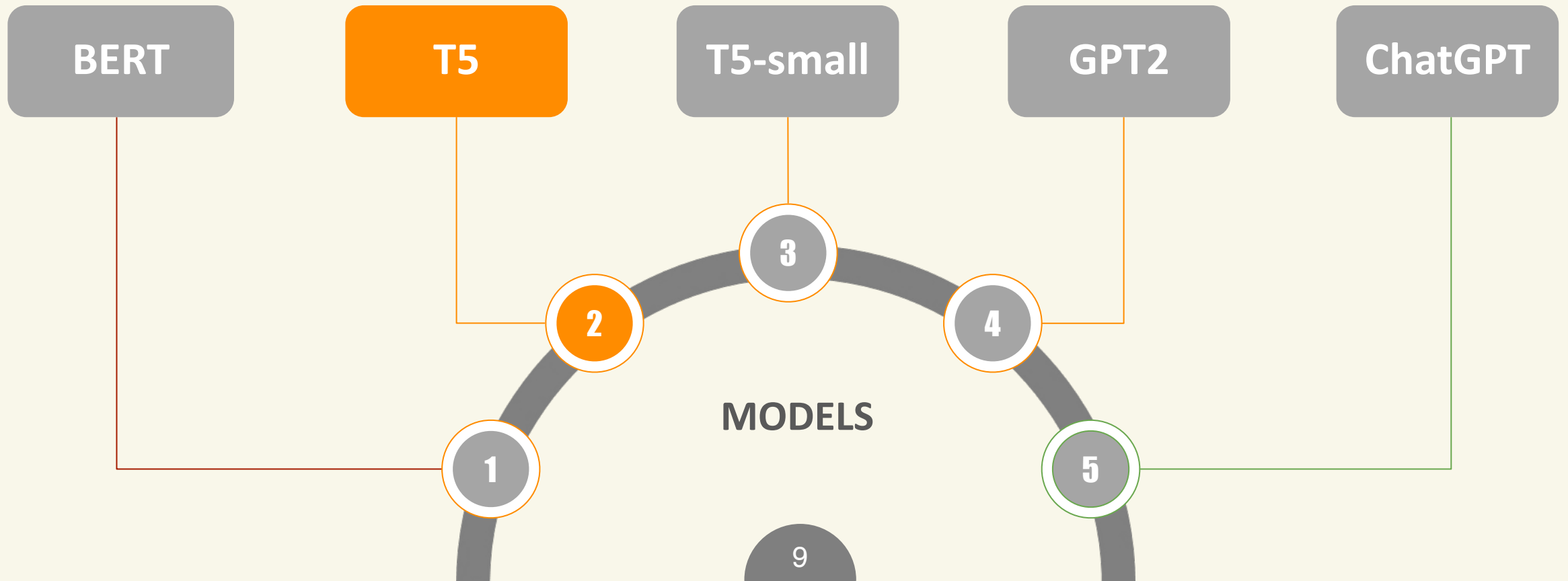
Fine-tune

Fine-tune
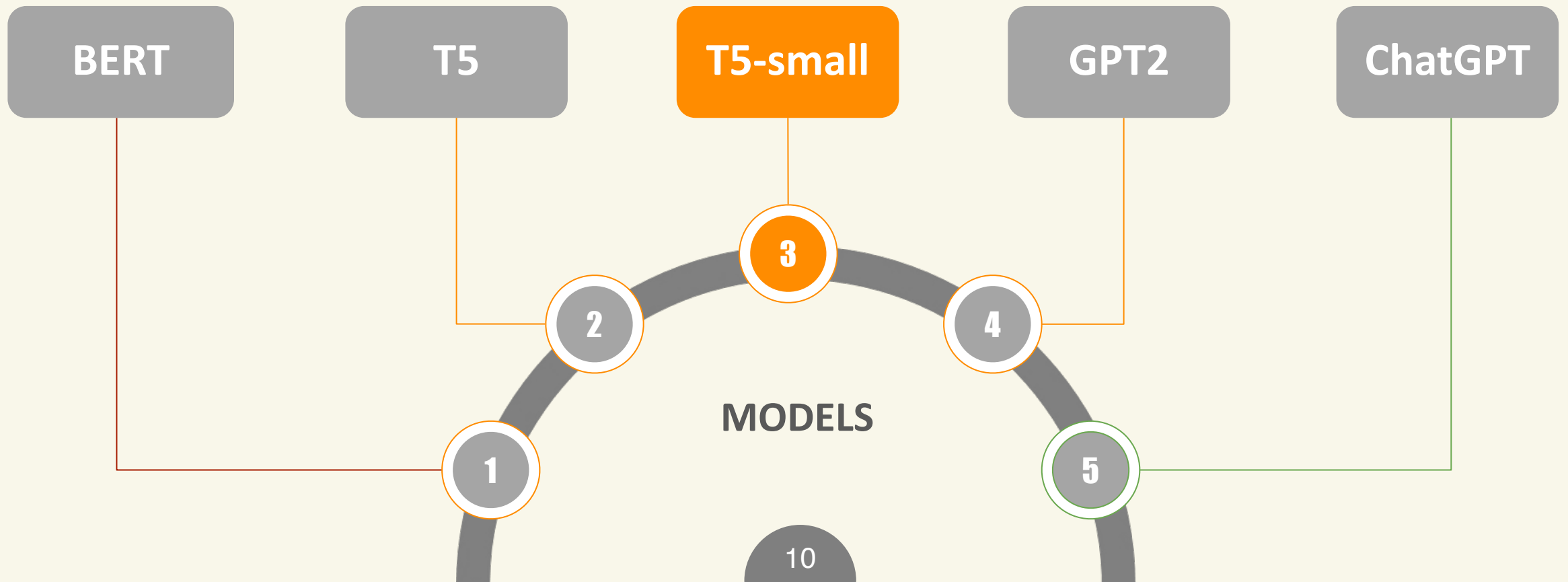&
Prompt Engineering

Prompt
Engineering
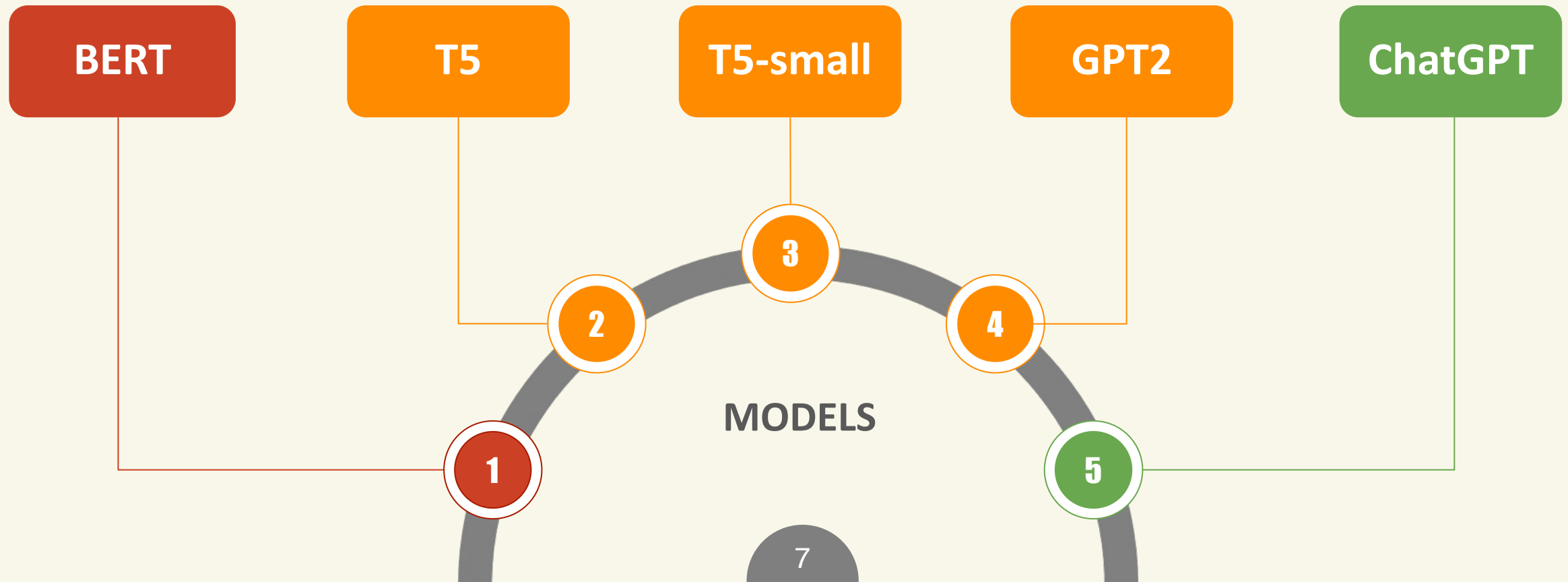
BERT

T5

T5-small

GPT2

ChatGPT

3

2

4

MODELS

1

5

7
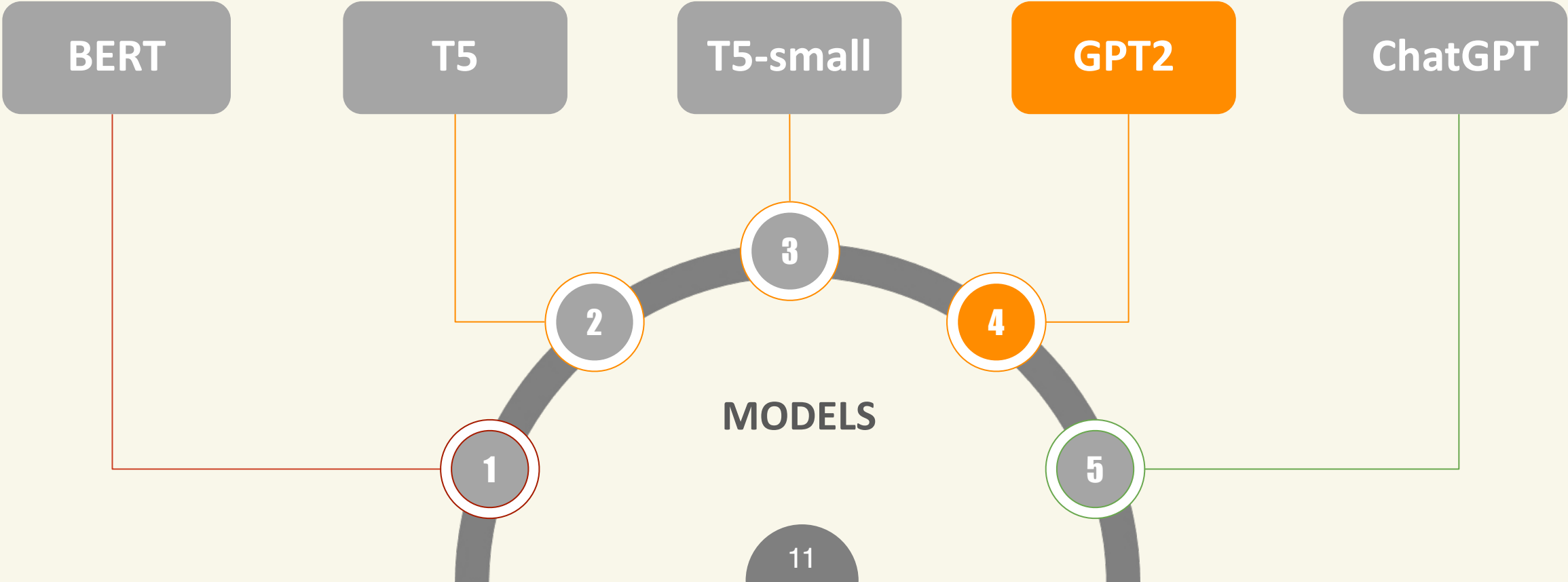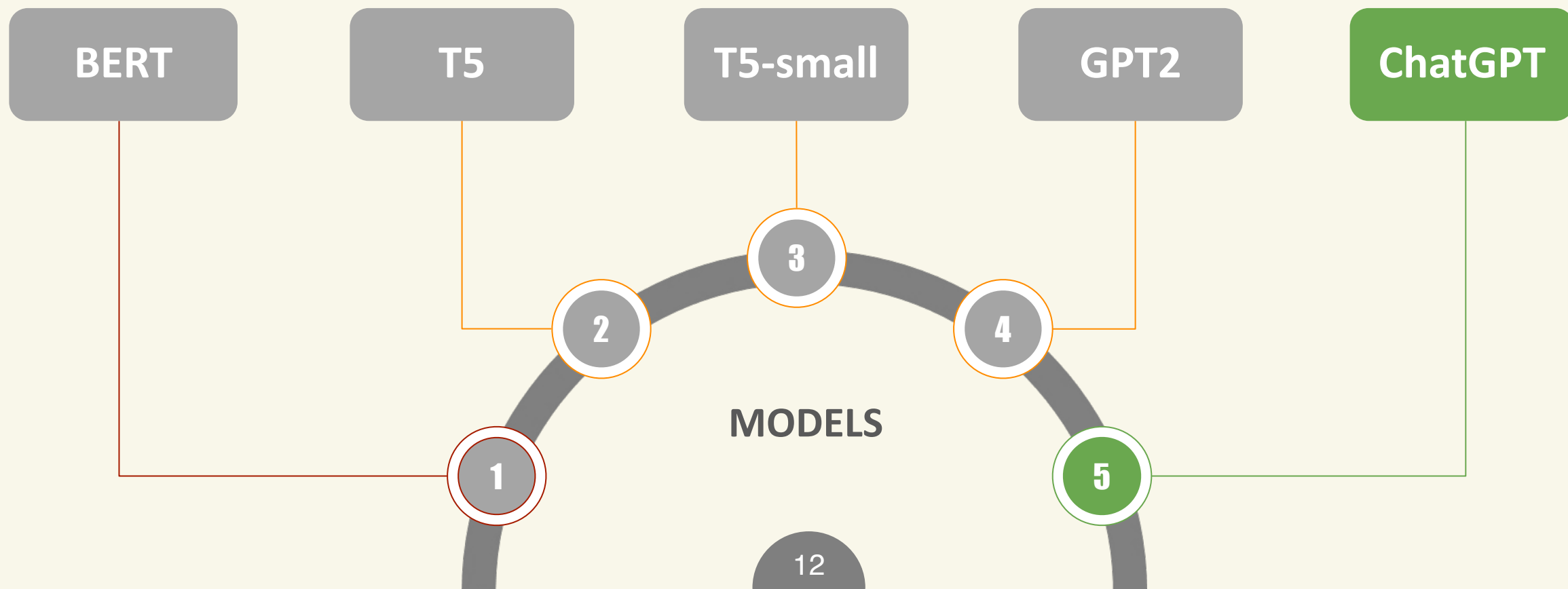
Fine-tuning

```
Train Step: 1100 Loss: 0.1759845021367073
Train Step: 1200 Loss: 0.16901450838893653
Train Step: 1300 Loss: 0.16636345073580074
Train Step: 1400 Loss: 0.17619289480149747
Train Step: 1500 Loss: 0.16598187167197465
Train Step: 1600 Loss: 0.17204287316650152
Train Step: 1700 Loss: 0.1782617973536253
Train Step: 1800 Loss: 0.1796263836324215
Train Step: 1900 Loss: 0.17243560660630464
Train Step: 2000 Loss: 0.20568211145699025
Train Step: 2100 Loss: 0.18708712719380854
Train Step: 2200 Loss: 0.1707417617738247
Train Step: 2300 Loss: 0.17130329988896847
Train Step: 2400 Loss: 0.19398574620485307
Epoch: 5 Loss is: 456.0205141287297
eval_loader len: 313
14842 13075
Epoch 5 Eval Acc: 0.8809459641557742; Eval Loss: 0.271360803145
5701
```

Prompt Engineering

```
dataset = ["""{
    "content": "世锦赛的整体水平远高于亚洲杯，要如同亚洲杯那样"鱼与熊掌兼得"，就需要各方面密切配合、#idiom#。作为主帅的俞觉敏，除了得打破保守思想，敢于破格用人，还得巧于用兵、#idiom#、灵活排阵，指挥得当，力争通过比赛推新人、出佳绩、出新的战斗力。",
    "realCount": 2,
    "groundTruth": ["通力合作", "有的放矢"],
    "candidates": [
        ["凭空捏造", "高头大马", "通力合作", "同舟共济", "和衷共济", "蓬头垢面", "紧锣密鼓"],
        ["叫苦连天", "量体裁衣", "金榜题名", "百战不殆", "知彼知己", "有的放矢", "风流才子"]
    ]
}"""]
```
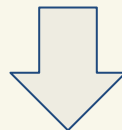
s = "世锦赛的整体水平远高于亚洲杯，要如同亚洲杯那样"鱼与熊掌兼得"，就需要各方面密切配合、通力合作。作为主帅的俞觉敏，除了得打破保守思想，敢于破格用人，还得巧于用兵、</s>、灵活排阵，指挥得当，力争通过比赛推新人、出佳绩、出新的战斗力。可选成语有：叫苦连天 | 量体裁衣 | 金榜题名 | 百战不殆 | 知彼知己 | 风流才子，请选出最佳的一项。"

		BERT	T5	T5-small	GPT2	ChatGPT
Fine-tuning	f1-score	0.545	0.454	0.41	0.644	–
	Accuracy	0.552	0.536	0.40	0.717	–
Prompt Engineering	f1-score	–	0.084	0	0.172	0.242
	Accuracy	–	0.148	0	0.198	0.246

Prompt engineering → fine-tuning

vs.

fine-tuning → prompt engineering

When we changed the data scale from 15,000 to **200,000**…

		BERT	**T5**	**T5-small**	**GPT2**	**ChatGPT**
Fine-tuning	f1-score	0.545	0.454	0.41	0.644	–
	Accuracy	0.552	0.536	0.40	0.717	–
Prompt Engineerin g	f1-score	–	0.084	0	0.172	0.242
	Accuracy	–	0.148	0	0.198	0.246

When we changed the data scale from 15,000 to **200,000**…

		BERT	T5	T5-small	GPT2	ChatGPT
Fine-tuning	f1-score	0.545	0.454	**0.70**	0.644	–
	Accuracy	0.552	0.536	**0.70**	0.717	–
Prompt Engineering	f1-score	–	0.084	0	0.172	0.242
	Accuracy	–	0.148	0	0.198	0.246

Fine-Tune	Prompt Engineering
- Discuss the scale of dataset and align the dataset and re-train, for the purpose of comparison - Compare the performance across different models	- Figure out the most appropriate on each of the model - Figure out the way to compare given different prompts

The End 🐝