# Calibrate Before Use: Improving Few-shot Performance of Language Models

https://arxiv.org/pdf/2102.09690.pdf

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, Sameer Singh. ICML 2021

**小林 颯介**
- Preferred Networks
- (東北大学 乾研 社会人D3 今月末で修了)

第13回 最先端NLP勉強会
2021年9月17日 オンライン

# 言語モデル Language Model; LM

- 文/文章の確率を与えるモデル
  - 単語列からその後に続く単語の確率 を計算するモデル (autoregressive LM)
  - 単語列の穴に当てはまる単語の確率 を… (masked LM) (※言語モデルじゃないと怒られることも)
- 昔〜: 音声認識・かな漢字変換・機械翻訳などに役に立つ
- ちょい昔1〜: DNN芸の~~サンドバッグ~~ベンチマークで流行った
- ちょい昔2〜: 良いエンコーダ(&デコーダ)パラメータの初期値になる
- 近年〜: 言語モデル自体がNLPタスクのソルバーとして使える

# 大規模言語モデルの圧倒的パワー

- プロンプト (prompt) による few-shot での解法
  - 連結のテキスト形式で Q1, A1, Q2, A2, …, 本番Q → [?] として入力し、出力単語として確率の高いものを解とする
  - 文途中の穴埋め単語予測で確率を比較することもある

Input: Subpar acting. Sentiment: Negative

Input: Beautiful film. Sentiment: Positive

Input: Amazing. Sentiment:

P(Negative|…) < P(Positive|…)

# プロンプト式

(few-shot in-context learning)

- 利点
  - 人々が即座にモデルを作れる・試せる
  - 誰でも使える; 機械学習モデルに自然言語インターフェースを付与
  - メモリ, 保存, システム複雑性の削減;タスク非依存1モデル
- 今回注目する欠点
  - プロンプト次第で性能が大きく変わってしまう
    (かつ、few-shot 設定では validation も難しい)

# プロンプトの3要素

1. フォーマット: プロンプトとラベルの “言葉選び”, テンプレート
2. 事例(集合): 例示的に見せる事例 (文 + ラベル)
3. 事例の順番: (複数の事例を)どの順番でつなげるか

Review: This movie is amazing!
Answer: Positive

Review: Horrific movie, don’t see it.
Answer:

Positive, Negative

Question: Did the author of the following tweet think that the movie was good or bad?
Tweet: This movie is amazing!
Answer: good

Question: Did the author of the following tweet think that the movie was good or bad?
Tweet: Horrific movie, don’t see it
Answer:

good, bad

# フォーマット・事例集合 は当然クリティカル

- [左右] フォーマット由来の性能変化
- [上下] 事例集合由来の性能変化
  (事例数 = 4)
- ちなみに最低のフォーマット(10)はこれ↓

This movie is amazing!
My overall feeling was that the movie was good

Horrific movie, don't see it.
My overall feeling was that the movie was

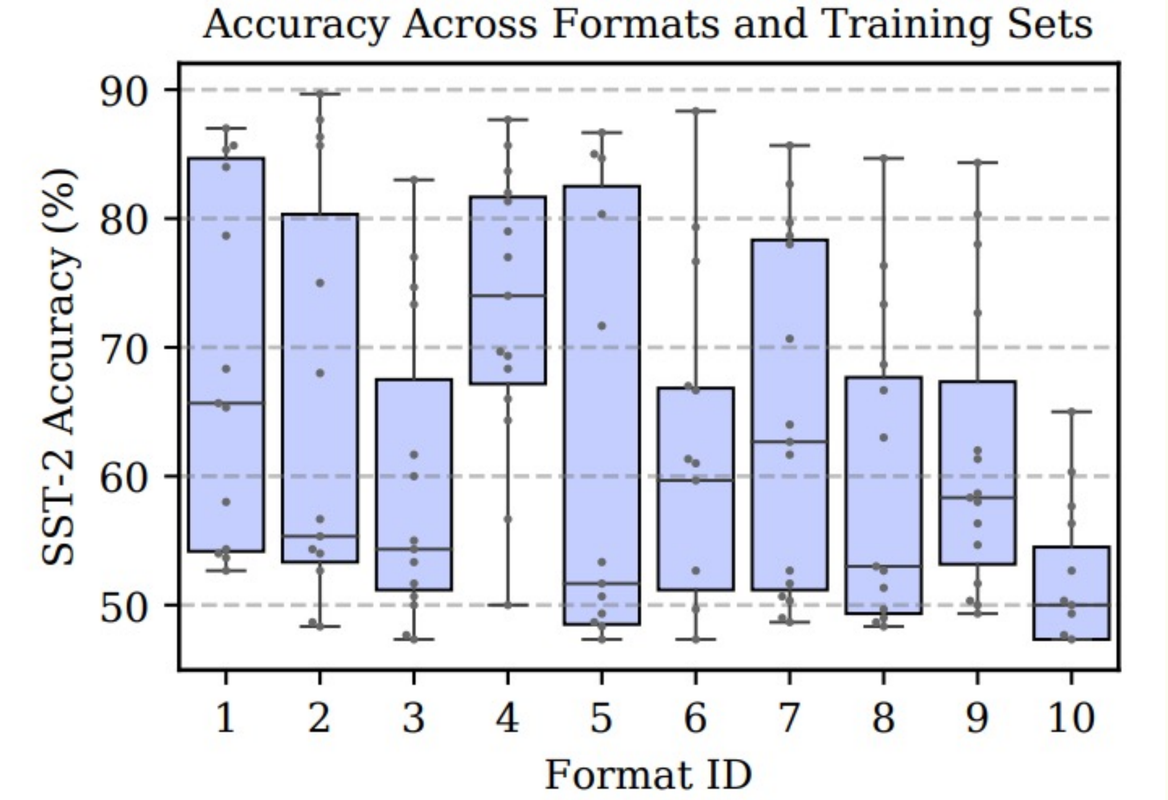



*Figure 3.* There is high variance in GPT-3's accuracy as we change the **prompt format**. In this figure, we use ten different prompt formats for SST-2. For each format, we plot GPT-3 2.7B's accuracy for different sets of four training examples, along with the quartiles.

# が、同じ事例集合でも順番がクリティカル

- [左右] 事例集合由来の性能変化
- [上下] 順番由来での性能変化

- 順番のインパクトがでかすぎる
  本来は非本質な情報なのに…
- こんな具合にどうもうさんくさい
  挙動なので調査&修正しよう！

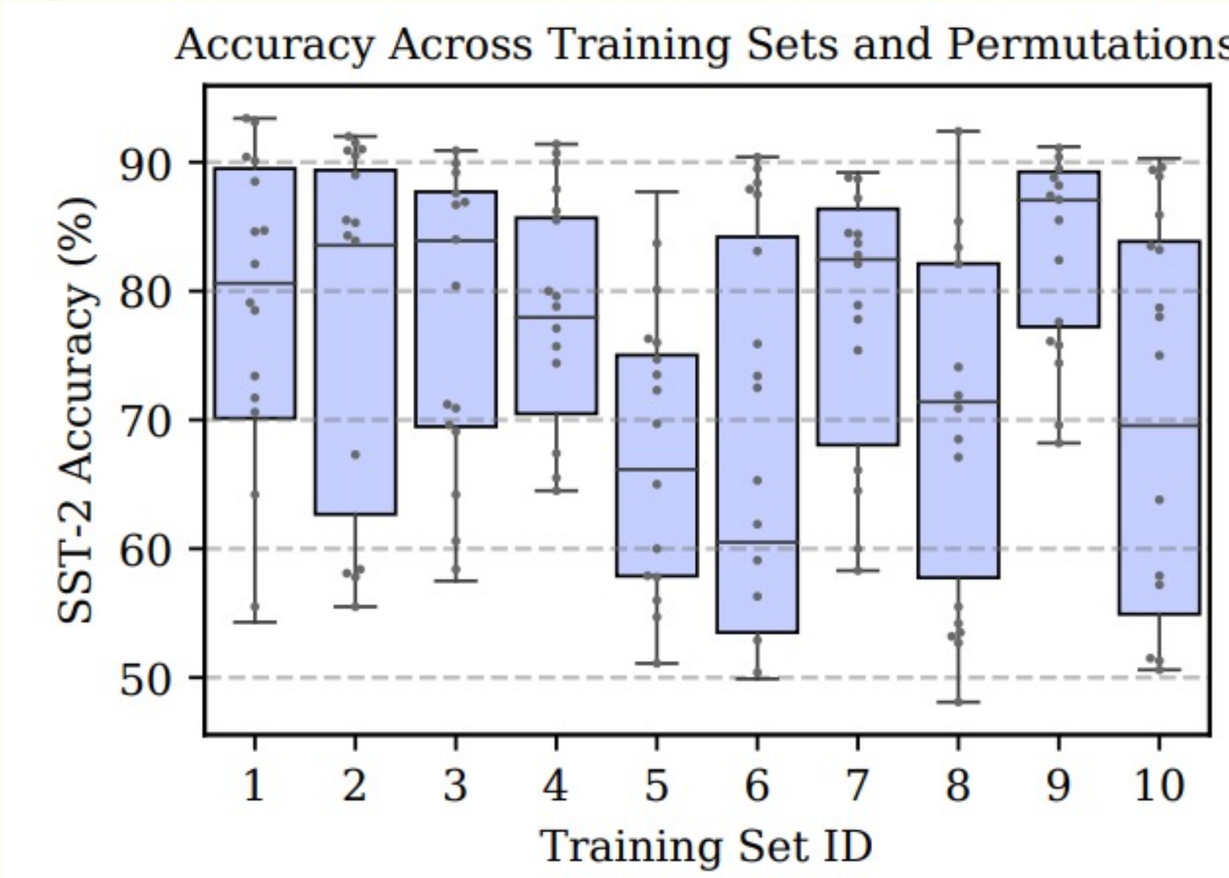
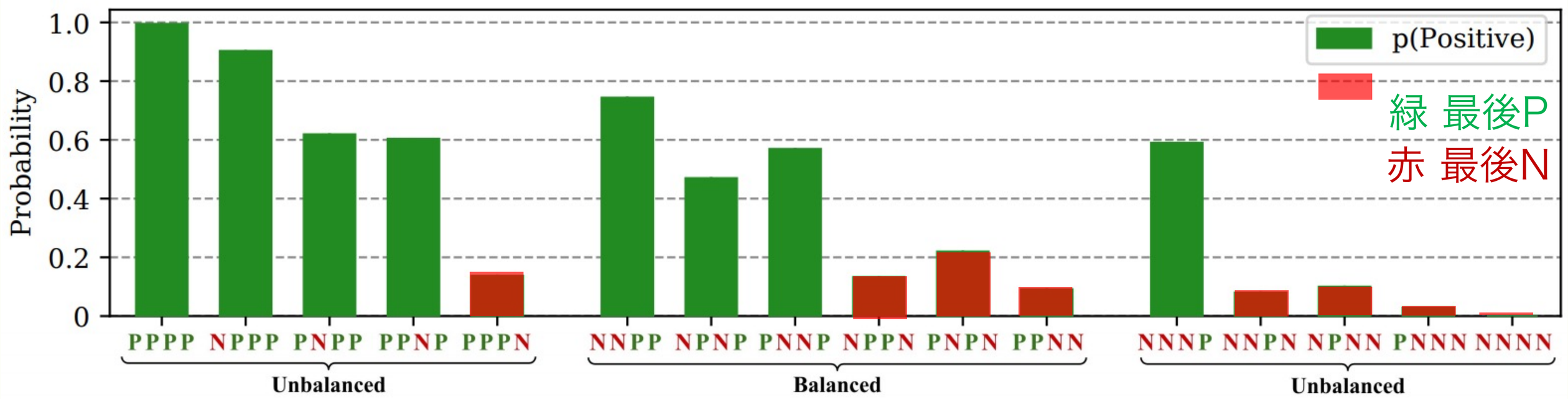



*Figure 2.* There is high variance in GPT-3's accuracy as we chan
the prompt's **training examples**, as well as the **permutation** of t
examples. Here, we select ten different sets of four SST-2 traini
examples. For each set of examples, we vary their permutation a
plot GPT-3 2.7B's accuracy for each permutation (and its quartile

# 予測にバイアスが乗る




- Posi/Nega分類. 4事例プロンプト後のPosi予測確率の平均値を比較
- 後のほうに見せた事例のラベルに引っ張られる; Recency bias
- 見せた事例のラベル頻度にも多少引っ張られる; Majority label bias
- (あと、素朴な単語頻度にも引っ張られる; Common token bias)

※普段、機械学習で Calibration といったら
『予測pが正解率p'に沿うようにする』などの狭義を指すことも多いので注意

# 簡単で効果的な対策: Contextual Calibration

- 『こんな入力が入ってきたとしても答えようがないだろう』
  という疑似事例 (context-free input) をデザインして
  『それに対して各ラベル確率が一様になるような補正』を求めて使う
- 実際に使った疑似事例: N/A [MASK] (入力文なし)
  (これらの補正の平均を使った)
- 補正方法 → 分類タスク: 割る補正 (W)
  生成タスク: 引く補正 (b), 1単語目のみ補正

```
p_cf = model("Review: This movie is amazing!
              Answer: Positive
              Review: N/A  Answer: ")
calibrated_model = lambda x: softmax(model(x) / p_cf)
# calibrated_model = lambda x: softmax(model(x) - p_cf)
```

$\hat{\mathbf{p}}$ :元のモデルが出す確率

補正後 $\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b})$

$\mathbf{W} = \text{diag}(\hat{\mathbf{p}}_{\text{cf}})^{-1}$

$\mathbf{b} = 0$

# 結果: 性能良くなるし安定もする

- [青] 改善. また、事例差やフォーマット差の分散も抑えられる

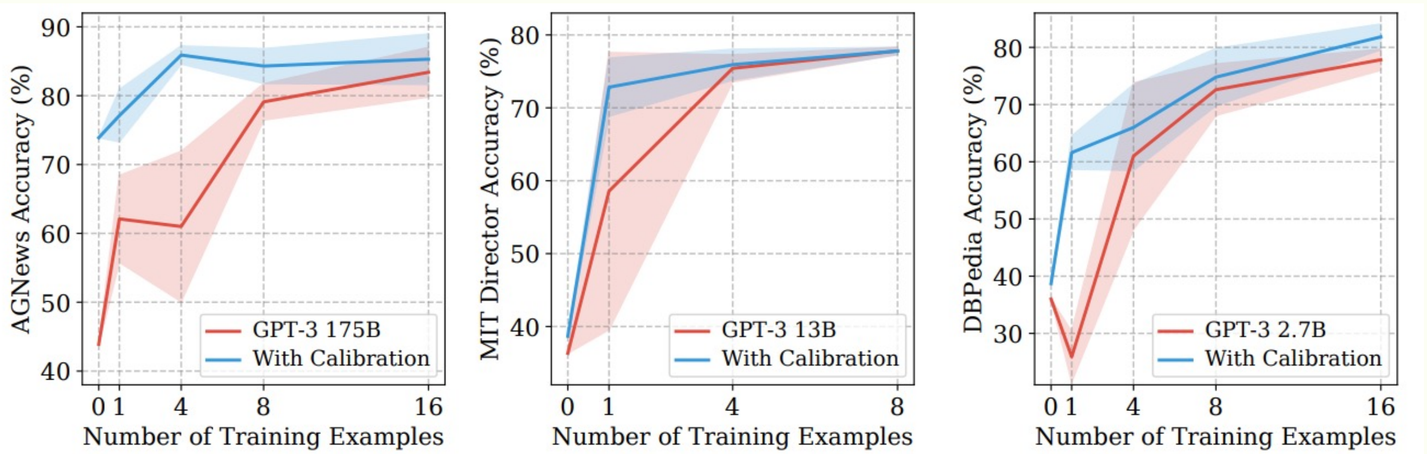



*Figure 1.* Few-shot learning can be highly unstable across different choices of the prompt. Above, we plot the mean accuracy (± one standard deviation) across different choices of the training examples for three different datasets and model sizes. We show that our method, *contextual calibration*, improves accuracy, reduces variance, and overall makes tools like GPT-3 more effective for end users.

↑[分散] 与える事例集合由来

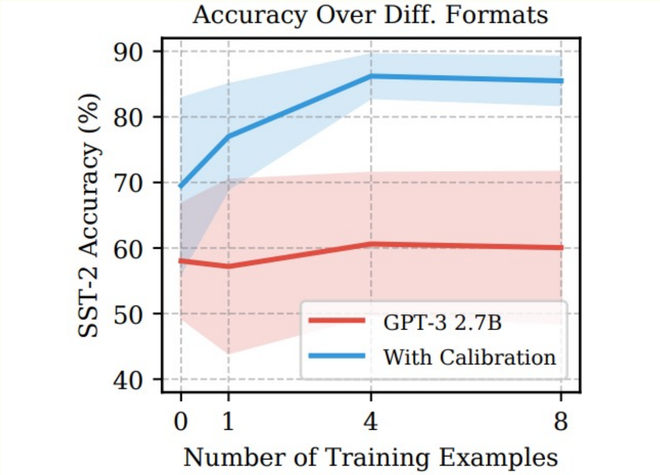



*Figure 7.* GPT-3 has high variance across different prompt formats; contextual calibration reduces this variance and improves mean accuracy. We show the mean accuracy (± standard deviation) over 15 different prompt formats for SST-2.

↑[分散] フォーマット由来

# どれくらい良い calibration?

- 一見雑な calibration
- だけどとても良い
- [緑] validation setで性能を最大化するようなWを計算して使った場合
  [青] 今回の雑 calibration
  [赤] 何もしないベースライン

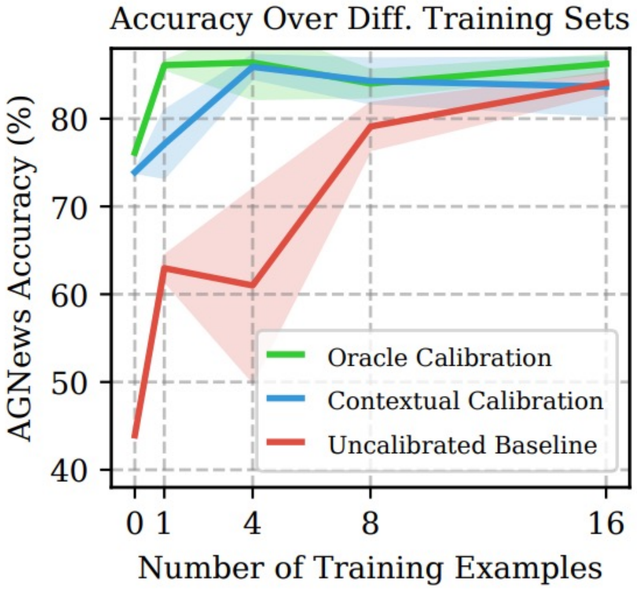



*Figure 8.* Contextual calibration, despite using no training data, achieves similar accuracy to an "oracle" calibration that finds the best $\mathbf{W}$ using the validation set. The plot shows GPT-3 175B's mean accuracy (± standard deviation) on AGNews over different choices of the training examples.

# calibration 用疑似事例 は雑でいい

- 一見雑な calibration
- 疑似事例は雑でいい. 他のでもいい
  - the, abc, the man., dasjhasjkdhjskdhds, nfjkhdvy84tr9bpuivwe

| Content-free Input | SST-2 | AGNews |
|---|---|---|
| Uncalibrated Baseline | 66.5 | 48.5 |
| N/A | 74.2 | 64.5 |
| [MASK] | 74.5 | 63.8 |
| ‘’ | 72.9 | 64.7 |
| N/A, [MASK], ‘’ | 79.0 | 66.5 |
| the | 69.1 | 59.0 |
| abc | 77.5 | 57.3 |
| the man. | 79.4 | 62.0 |
| dasjhasjkdhjskdhds | 79.3 | 64.5 |
| nfjkhdvy84tr9bpuirvwe | 78.4 | 65.5 |

*Table 3.* We show the accuracy for 1-shot SST-2 and 0-shot AGNews over different choices for the content-free input. The choice of content-free input matters, however, *many good choices exist*. The token ‘’ indicates the empty string. Recall that in our experiments, we ensemble over N/A, [MASK], and the empty string.

# まとめ

- プロンプトで性能大きく変わる問題 の中でも3つの現象をピックアップ
  - Majority label bias: プロンプト中の事例のラベル頻度に引っ張られる
  - Recency bias: プロンプト終盤の事例のラベルに引っ張られる
  - Common token bias: 言語モデル学習時の単語頻度に引っ張られる
- Contextual calibration という簡単な事前補正処理を提案
  - 「N/A: → 全ラベルに対して一様な確率を出す」を満たす scaling or bias
- 最大30%改善. プロンプトでの性能分散も抑える
  - prompt engineering の手間削減

# Appendix

- 似た問題意識だけど「良い順番を探す」アプローチの論文
  - Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity https://arxiv.org/abs/2104.08786
- Promptとか指示系の文に教師あり学習で慣らせばいい
  - Finetuned Language Models Are Zero-Shot Learners https://arxiv.org/abs/2109.01652
- Promptサーベイ
  - Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing https://arxiv.org/abs/2107.13586

# Appendix

| Task | Prompt |
|---|---|
| LAMA | Alexander Berntsson was born in Sweden<br><br>Khalid Karami was born in |
| ATIS<br>(Airline) | Sentence: what are the two american airlines flights that leave from dallas to san francisco in the evening<br>Airline name: american airlines<br><br>Sentence: list a flight on american airlines from toronto to san diego<br>Airline name: |
| ATIS<br>(Depart Date) | Sentence: please list any flight available leaving oakland california tuesday arriving philadelphia wednesday<br>Depart date - Day name: tuesday<br><br>Sentence: show me all all flights from pittsburgh to atlanta on wednesday which leave before noon and serve breakfast<br>Depart date - Day name: |
| MIT Movies<br>(Genre) | Sentence: last to a famous series of animated movies about a big green ogre and his donkey and cat friends<br>Genre: animated<br><br>Sentence: what is a great comedy featuring the talents of steve carell as a loser looking for a friend<br>Genre: |
| MIT Movies<br>(Director) | Sentence: in 2005 director christopher nolan rebooted a legendary dc comics superhero with a darker grittier edge in which movie<br>Director: christopher nolan<br><br>Sentence: what 1967 mike nichols film features dustin hoffman in romantic interludes with anne bancroft as mrs robinson<br>Director: |

*Table 6.* The prompts used for generation tasks. We show one training example per task for illustration purposes.