Are Prompt-based Models Clueless?

Pride Kavumba[1,2] **Ryo Takahashi**[3] **Yasuke Oda**[3,1]
[1]Tohoku University [2]RIKEN AIP [3]LegalForce Research
kavumba.pride.q2@dc.tohoku.ac.jp
{ryo.takahashi, yusuke.oda}@legalforce.co.jp

Abstract

Finetuning large pre-trained language models with a task-specific head has advanced the state-of-the-art on many natural language understanding benchmarks. However, models with a task-specific head require a lot of training data, making them susceptible to learning and exploiting dataset-specific superficial cues that do not generalize to other datasets. Prompting has reduced the data requirement by reusing the language model head and formatting the task input to match the pre-training objective. Therefore, it is expected that few-shot prompt-based models do not exploit superficial cues. This paper presents an empirical examination of whether few-shot prompt-based models also exploit superficial cues. Analyzing few-shot prompt-based models on MNLI, SNLI, HANS, and COPA has revealed that prompt-based models also exploit superficial cues. While the models perform well on instances with superficial cues, they often underperform or only marginally outperform random accuracy on instances without superficial cues.

1 Introduction

Finetuning large pre-trained language models with a task-specific head has achieved remarkable performance on many natural language benchmarks (Wang et al., 2018, 2019). However, the task-specific head introduces a lot of random task-specific parameters that require enormous finetuning data to attain optimal performance. The exposure to enormous data increases the potential for models to learn and exploit dataset-specific superficial cues that do not generalize to other datasets without superficial cues (Gururangan et al., 2018; Poliak et al., 2018; Sugawara et al., 2018; Niven and Kao, 2019; Schuster et al., 2019; Kavumba et al., 2019). For example, Niven and Kao (2019) found that task-specific head models exploit the presence of "not" in the input of argument reasoning comprehension dataset (Habernal et al., 2018)

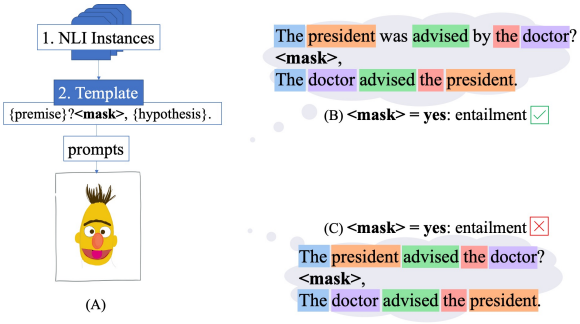


Figure 1: (A) shows a prompt-based model receiving natural language inference (NLI) prompts generated through a template. (B) Our analysis reveals that prompt-based models exploit superficial cues (highlighted lexical overlap) that are predictive of entailment relation between premise and hypothesis. (C) The model fails to generalize to instances were the superficial cues no longer predict entailment relation.

to achieve state-of-the-art accuracy, but drop to random accuracy when the superficial cue is neutralized. On the other hand, prompting reuses the pre-training language model head, introducing no random task-specific parameters. Thus, prompt-based models can achieve remarkable performance with only a few training examples (Brown et al., 2020; Schick and Schütze, 2021a,b; Gao et al., 2021; Le Scao and Rush, 2021). Hence, few-shot prompting lowers the potential for models to learn and exploit dataset-specific superficial cues.

This work empirically investigates whether few-shot prompt-based models exploit superficial cues. Specifically, we ask: *Do few-shot prompt-based models exploit superficial cues?* To answer this question, we examine prompted-based models on two fundamental tasks of natural language understanding: natural language inference (NLI) and commonsense reasoning; comprehending natural language inference and commonsense are essential to make progress in natural language understanding (Bowman et al., 2015; Williams et al., 2018; Roemmele et al., 2011). We analyze the per-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2333 - 2352

formance of prompt-based models trained on the Stanford Natural Language Inference dataset (Bowman et al., 2015, SNLI), the Multi-Genre Natural Language Inference data (Williams et al., 2018, MNLI), and the Choice of Plausible Alternatives dataset (Roemmele et al., 2011, COPA) on instances with and without superficial cues.

To facilitate the analysis, we define two types of superficial cues that abstract away from the underlying tasks: *context* and *contextless* superficial cues, where the definition of the *context* depends on the task. For example, in natural language inference tasks, we define the premise as the *context*, while in multiple-choice tasks, we define the question as the *context*. Context superficial cues such as lexical overlap (Figure 1) coexist in the context (premise) and the hypothesis. In contrast, contextless superficial cues exist only in the hypothesis (in NLI) or in answer choices (in multiple-choice tasks). A dataset can contain either one or both types of superficial cues. Therefore, both types must be investigated to sufficiently answer whether a model can exploit superficial cues.

As a prerequisite, we reanalyze superficial cues in MNLI, SNLI, and COPA datasets to created evaluation sets that have and do not have superficial cues. We find that these datasets contain more superficial cues than previously known. Specifically, we find that 90.1% of MNLI matched instances contain *contextless superficial cues* in the hypothesis, while 71.9% SNLI contains *contextless superficial cues*. Additionally, we find that COPA contains not only contextless superficial cues (Kavumba et al., 2019), but 78.0% of the instances also contain *context superficial cues*.

Finally, we examine whether few-shot prompt-based models also rely on superficial cues to achieve remarkable performance on MNLI, SNLI, COPA, and the HANS dataset (McCoy et al., 2019). COPA experiments reveal that prompt-based models do not rely on contextless superficial cues for typical few-shot training sizes. However, the other empirical results show that prompt-based models heavily rely on superficial cues—failing to generalize to data without superficial cues (Figure 1).

In summary, our contributions are:

1. We propose to divide superficial cues into *context* and *contextless* superficial cues, which abstracts away from the underlying tasks (§ 3).
2. We established that the datasets of MNLI (§ 3.2), SNLI (§ 3.2) and COPA (§ 3.1) contain more superficial cues than previously known. We release analyzed datasets at `https://github.com/legalforce-research/prompt-models-clueless`.
3. We present the first investigation of the exploitation of superficial cues by prompt-based models, finding that prompt-based models also exploit superficial cues (§ 5).

2 Background

We will begin with a review of the necessary concepts required to understand the rest of the paper.

2.1 Prompting

Prompt-based finetuning has been demonstrated to be effective in few-shot setup (Brown et al., 2020; Schick and Schütze, 2021a,b; Gao et al., 2021; Le Scao and Rush, 2021). By reusing the pretraining language model head, prompting introduces no or only a few randomly initialized parameters. Prompting reformulates any task to match the pretraining objective. For example, consider the task of classifying the sentiment polarity of movie reviews using a masked language model such as BERT (Devlin et al., 2019). A review such as "*I liked the movie*" is converted to "*I liked the movie. It was [MASK]*". The model, then, fills [MASK] with words such as *{good, nice, bad, terrible}*, which are mapped to the task labels—positive or negative—through a verbalizer (Schick and Schütze, 2021a,b). In contrast, a task-specific classification head model directly predicts positive or negative sentiment. For an in-depth review, we will refer the interested reader to a survey by Liu et al. (2021).

2.2 Superficial Cues

Superficial cues can be described as linguistic or non-linguistic characteristics of instances that have nothing to do with the task itself but are tied to a specific task label. These characteristics include lexical overlap (McCoy et al., 2019), distinct words frequently appearing in the correct choices (Niven and Kao, 2019; Kavumba et al., 2019), and distinctive style of the correct choices (Trichelair et al., 2019). As a concrete example, consider a sentiment classification dataset whose negative sentiment instances contain "not"; for example, "I did not like the movie". Here, "not" is a superficial cue because it is predictive of the correct label.

2.3 Training Datasets

MNLI The Multi-Genre Natural Language Inference (Williams et al., 2018, MNLI) dataset is an important dataset of natural language inference which is also part of the SuperGLUE benchmark (Wang et al., 2019). Given a premise and a hypothesis, the task asks to pick one label from among three, {contradiction, neutral, entailment}. The test set of MNLI is divided into matched (in-domain instances) and mismatched (out-of-domain instances) subsets based on whether the domain of each test instance matches the training set domain.

SNLI The Stanford Natural Language Inference (Bowman et al., 2015, SNLI) is a popular natural language inference dataset with the same format as MNLI.

COPA The Choice of Plausible Alternatives (Roemmele et al., 2011, COPA) dataset is a popular multiple-choice commonsense dataset, which is also a part of the SuperGLUE benchmark. Given a premise and a question, the task is to select the most plausible cause or effect from the set of two candidates.

3 Superficial Cues in NLI and COPA

We investigate prompted-based models on two fundamental tasks of natural language understanding: natural language inference (NLI) and commonsense reasoning. As a prerequisite, we begin by creating test sets with and without superficial cues that we will subsequently use to investigate whether prompt-based models exploit superficial cues. We analyze and split test sets of English language datasets into subsets with and without superficial cues in the following subsections.

To facilitate easy analysis, we divide superficial cues into two categories: *context* superficial cues and *contextless* superficial cues, where the definition of *context* is task dependent. For example, in natural language inference tasks such as COPA, the *context* can be defined as the premise, while in multiple-choice tasks, the *context* can be defined as the question. Context superficial cues, such as lexical overlap found by McCoy et al. (2019) can only be exploited when the context is available in the input. On the other hand, contextless superficial cues, such as the occurrence of "not" in the correct answer choices found by Niven and Kao (2019), are those that are exploitable even in the absence of the context required to perform a task.

3.1 Context Superficial Cues

Natural Language Inference (NLI): NLI has a good dataset designed to test for contextless superficial cues. Specifically, the HANS dataset tests the models' ability to exploit three types of context superficial cues in NLI: *lexical overlap*, *subsequence*, and *constituent* McCoy et al. (2019). Therefore, we evaluate prompt-based models on the HANS dataset instead of splitting tests of MNLI and SNLI into instances with and without superficial cues.

COPA Eyeballing all instances to find common patterns that identify the correct answer choice, but are unrelated to the task, can be challenging and error-prone. To circumvent the need for manual examination, we propose to solve the task in a setup that encourages the model to solve the task using superficial cues. This setup is similar to providing only partial input (Gururangan et al., 2018; Poliak et al., 2018). Specifically, we randomly shuffle the words in the answer choices such that identifying the correct choice is mainly based on superficial cues in the question and the answer choice. For example, given the original instance;

Premise: The host cancelled the party. What was the CAUSE of this?
a) She worried she would catch the flu.
b) She was certain she had the flu. (correct)

The new answer choices for the new instance becomes:

a) She would she catch the worried flu.
b) She had was she the certain flu. (correct)

In this setting, we find that RoBERTa achieves an average accuracy of 78%, indicating the existence of context superficial cues. Following this result, we split the test set into a subset with superficial cues containing instances solved by the majority of models, and a subset without superficial cues, containing all the remaining instances.

3.2 Contextless Superficial Cues

Natural Language Inference To investigate contextless superficial cues in NLI, we train RoBERTa (Liu et al., 2019) with a classification head on only the hypothesis of MNLI and SNLI. This analysis is similar to the one done by Gururangan et al. (2018) using fastText (Joulin et al., 2017). If the model can not find contextless superficial cues in the hypothesis, it is expected to achieve random performance (33.3%). But, RoBERTa trained

Dataset	accuracy
Random	33.3
MNLI	90.1 ±0.1
MNLI-mm	90.0 ±0.2
SNLI	71.9 ±0.1

Table 1: Average accuracy on matched MNLI (MNLI) and mismatched MNLI (MNLI-mm), and SNLI for a head RoBERTa model trained on the hypothesis. MNLI and MNLI-mm results are for a model trained on MNLI and SNLI results are for a model trained on SNLI.

on MNLI achieve an average performance of 90.1% and 90.0% on matched and mismatched instances, respectively (Table 1), which is worse than previously known (53.9% matched and 52.3% mismatched (Gururangan et al., 2018)). On the test set of SNLI, RoBERTa trained on SNLI achieves an average accuracy of 71.9% (Table 1), which is 4.9 percentage points higher than previously known Gururangan et al. (2018). Following this result, we split the testing sets of MNLI and SNLI such that each test set has two subsets: instances with contextless superficial cues, containing all instances that the majority of models solved correctly, and instances without contextless superficial cues contain all the remaining instances.

COPA The test set of COPA has already been split into two subsets that have instances with contextless superficial cues and instances that do not have contextless superficial cues Kavumba et al. (2019). The subsets were constructed based on the performance of RoBERTa trained on answers only. Therefore we do not reanalyze COPA; instead, we will use the same publicly available subsets in our evaluation.

4 Experimental Setup

The goal of our experimental setup is to answer the following research question: Do prompt-based models exploit superficial cues? We decompose this question into two sub-questions: 1) Do prompt-based models exploit context superficial cues? 2) Do prompt-based models exploit contextless superficial cues?

Training Details For all our experiments, we use RoBERTa-large (355M parameters) because it is the widely used model in prompt-based fine-tuning (Schick and Schütze, 2021a; Gao et al., 2021; Le Scao and Rush, 2021). We build on the source code by Gao et al. (2021) [1] and Le Scao and Rush (2021) [2], and we load the pre-trained weights from HuggingFace (Wolf et al., 2019). We use the best-reported hyperparameters and templates (Appendix B) from Gao et al. (2021) on NLI. All NLI models are trained with 16 instances per label. We use the same partitions used by Gao et al. (2021). On COPA we use the best hyperparameters and templates (Appendix B) from Schick and Schütze (2021b); Le Scao and Rush (2021). We ran all experiments three times with different random seeds and report the average and standard deviation.

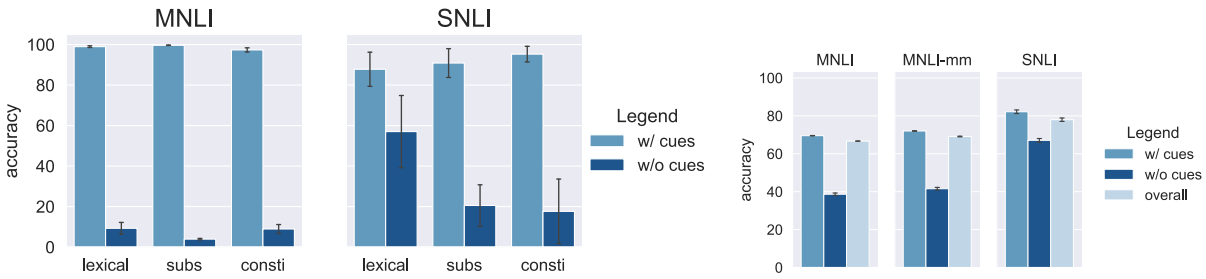
Evaluation The goal of our experimental evaluation is to answer the following question: Do prompt-based models exploit superficial cues? We answer this question by investigating whether the model exploits or relies on either context or contextless superficial cues. We train and evaluate our models on English datasets.

Context Superficial Cues To investigate whether models exploit context superficial cues, we train prompt-based models on MNLI, SNLI, and COPA. We evaluate NLI models on the HANS dataset that tests the models' ability to exploit three types of context superficial cues in NLI: *lexical overlap*, *subsequence*, and *constituent* McCoy et al. (2019). We report the average accuracy and standard deviation on the two subsets of the dataset: a subset where the superficial is informative (Entailment) and a subset where the superficial cues are uninformative (Non-entailment). A model that does not rely on context superficial cues is expected to perform comparably on both subsets.

Contextless Superficial Cues To investigate whether prompt-based models exploit contextless superficial cues, we train a prompt-based model on MNLI, SNLI, and COPA; and evaluate them on the corresponding test set of each dataset. Each test set consists of two subsets obtained and described in sections 3: a subset of instances with contextless superficial cues and a subset without contextless superficial cues. A model that does not rely on contextless superficial cues is expected to perform comparably on both subsets.

[1] https://github.com/princeton-nlp/LM-BFF

[2] https://github.com/TevenLeScao/pet



(a) w/ & w/o Context Cues

(b) w/ & w/o Contextless Cues

Figure 2: Average accuracy on instances with superficial cues (w/ cues) and without superficial cues (w/o cues) for prompt-based RoBERTa models trained on MNLI and SNLI. (a) shows results on the HANS dataset on three kinds of context superficial cues: lexical overlap (lexical), subsequence (subs), and constituent (consti). (b) shows results on instances with and without contextless superficial cues of matched MNLI (MNLI) and mismatched MNLI (MNLI-mm) for a model trained on MNLI; and evaluations results on SNLI for a model trained on SNLI.

5 Results

5.1 Exploiting Context Superficial Cues

Natural Language Inference (NLI) Figure 2a shows the results on the HANS dataset of the prompt-based model trained on MNLI (left) and SNLI (right), respectively. The results show that prompt-based RoBERTa trained on MNLI performs considerably well on instances with superficial cues, an overall average of 98.7%. However, the model only achieves an overall average accuracy of 7.4% on instances without context superficial cues, failing to reach random accuracy of 50%. This indicates that the prompt-based models trained on MNLI exploit superficial cues.

Similarly, figure 2a (right) shows that while RoBERTa performs considerably well on instances with superficial cues (overall average 91.3%), it fails to achieve the same performance on instances without superficial cues (overall average of 31.7%). [3] This result also leads to the same conclusion that the model exploits context superficial cues.

COPA Table 2 shows the results of the prompt-based RoBERTa trained on COPA and evaluated on the two subsets of COPA: with and without superficial cues. The results show RoBERTa performs well on instances with superficial cues but barely exceeds random accuracy (50%) on instances without superficial cues. This, too, indicates that the model exploit contextless superficial cues.

[3]The high variance is similar to that reported by previous work studying head models (Bras et al., 2020).

# Examples	*w/ cues*	*w/o cues*	Overall
8	83.0±0.5	55.7±1.4	78.1±0.3
16	81.6±1.9	57.9±1.0	77.3±1.4
32	82.4±2.0	53.5±0.5	77.1±1.5
64	84.4±1.4	53.8±4.1	78.9±1.2
96	87.0±1.7	57.9±4.1	81.7±0.8
100	87.9±1.7	54.2±1.9	81.7±1.2

Table 2: Average accuracy of few-shot prompt-based RoBERTa models on COPA instances with context superficial cues between the context and the answer choices; and instances without exploitable context superficial cues. Column "# Examples" shows the number of examples used for finetuning.

5.2 Exploiting Contextless Superficial Cues

NLI Figure 2b shows the results of prompt-based RoBERTa train on MNLI and SNLI and evaluated on the corresponding test set. The results show that the prompt-based model trained on MNLI performs considerably better on instances with superficial cues on both matched (69.5%) and mismatched (72.0%) instances. On instances without superficial cues we observe a gap of 30.9% and 30.5% on matched and mismatched instances, respectively. The high difference in performance indicates that the models exploit contextless superficial cues. For the model trained on SNLI and evaluated on SNLI subsets, we observe a gap of 15.1% between performance on instances with and without superficial cues (82.2% vs 67.1%). This also indicates that the model does exploit contextless superficial cues.

COPA Table 3 shows of the prompt-based RoBERTa trained on COPA and evaluated on the

# Examples	*w/ cues*	*w/o cues*	Overall
8	79.1±0.9	77.4±0.0	78.1±0.3
16	77.0±1.3	77.4±1.6	77.3±1.4
32	77.7±3.1	76.8±0.9	77.1±1.5
64	80.2±0.5	78.1±1.6	78.9±1.2
96	83.7±1.1	80.4±0.7	81.7±0.8
100	84.6±1.1	80.0±1.6	81.7±1.2

Table 3: Average accuracy of few-shot prompt-based RoBERTa models on COPA instances with contextless superficial cues (w/ cues) in the answer choices and instances without contextless superficial cues (w/o cues). Column "# Examples" shows the number of examples used for finetuning.

COPA subsets with and without superficial cues. The results show that the prompt-based model does not exploit superficial cues at a small enough training set (less or equal to 32 instances). However, increasing the size further increases the gap in performance between instances with and without contextless superficial cues. It is encouraging to note that the model does not exploit contextless superficial cues at sizes commonly used in few-shot settings.

6 Discussion

6.1 Predictions Errors

The results on natural language inference instances without superficial cues are worse than random performance. One wonders whether it is because the instances are hard. We look at some instances that the model fails to solve correctly. We show some of the instances in Table 4. The instances are simple enough for anyone that understands English. One question that immediately arises is; are prompt-based models sensitive to the meaning of the question?

6.2 Attention Visualization

To investigate whether prompt-based models are sensitive to meaning, we compare the attention weight across all twenty-four layers of RoBERTa-large for closely related instances that differ only in meaning and hence the labels. While there have been many questions that have been raised over the reliability of singly using attention weights for explanation (Wiegreffe and Pinter, 2019; Vig and Belinkov, 2019), here we use attention weights coupled with other results to gain more insights into the models' inner working. We are interested in knowing whether there is a huge change in attention weights responding to the change in meaning. For example, we take an instance with superficial cues, which lead to the correct prediction of the Entailment label:

Premise: The president was advised by the doctor.
Hypothesis: The doctor advised the president.
label: Entailment

And an instance without superficial cues:

Premise: The president advised the doctor.
Hypothesis: The doctor advised the president.
label: Non-Entailment

While the instances are completely different in meaning, the model predicts Entailment in both cases because of the superficial cue of high overlap. When we compare the attention maps for all the layers, we notice that there is barely any change in response to the change in the meaning of the sentences. Because of space limitation, we show attention maps only for the first two layers and the last layer (Figure 3). The attention maps for all the 24 layers are shown in Appendix E. The visualizations highlight the inability of the model to respond to the change in meaning. We investigate this further in the next section.

6.3 Sensitivity to Word Order in NLI

The visual attention analysis revealed that the model does not respond well to change in meaning. To investigate this at scale, we evaluate a model trained on input with correct word order and input with randomly shuffled word order. We will refer to input with correct word order as *meaningful* input and input with shuffled word order as *meaningless* input. Specifically, given an original test instance, we make it meaningless by randomly shuffling all the words in the instance while maintaining the English end of sentence punctuation mark if it exists in the original instance. We do this so we can preserve the same number of English sentences as the original instance while making them meaningless.

For example, given the original NLI instance:

Premise: The president was advised by the doctor.
Hypothesis: The doctor advised the president.
label: Entailment

The new instance becomes:

Premise: The doctor by the president advised was.

Premise	Hypothesis	Label	Prediction
The president advised the doctor	The doctor advised the president	N	E
The student saw the managers	The managers saw the student	N	E
The presidents encouraged the banker	The bank encouraged the president	N	E
The actors avoided the bankers	The bankers avoided the bankers	N	E
The managers saw the secretaries	The secretary saw the managers	N	E
The lawyers helped the judge	The judge helped the lawyers	N	E
The banker thanked the tourist	The tourist thanked the banker	N	E

Table 4: Some examples of instances without superficial cues from the HANS dataset that a prompt-based model trained on MNLI wrongly classify as entailment (E) because of superficial cues: high lexical overlap.

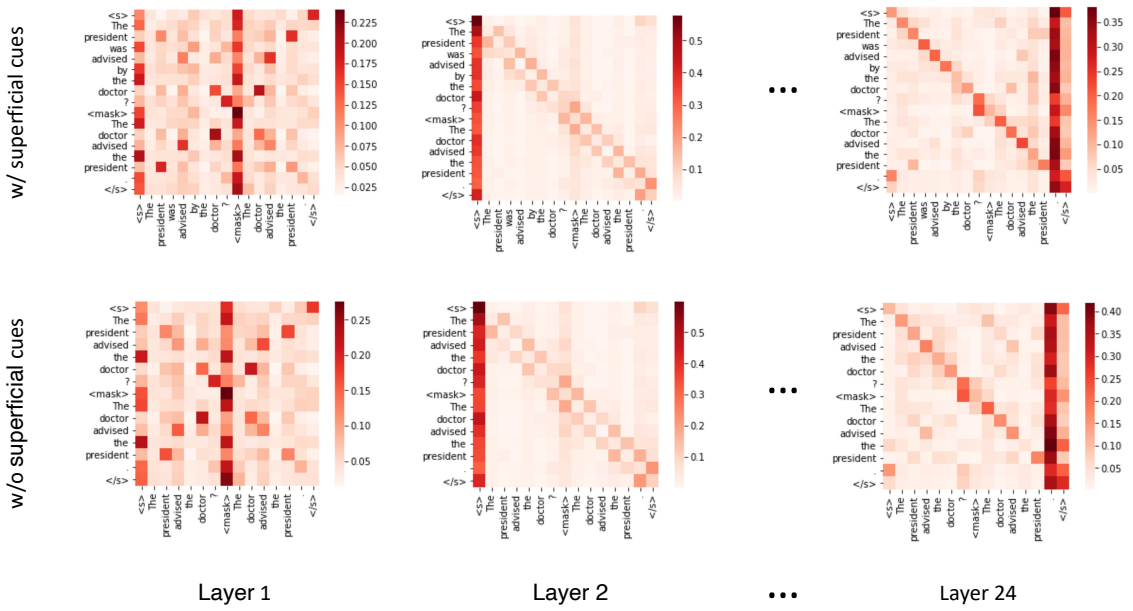


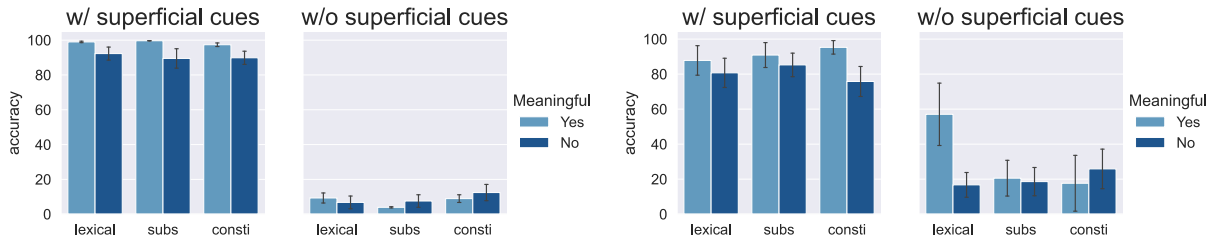
Figure 3: A comparison of prompt-based model attention maps on two natural language inference instances that have some words in common but have entirely different meanings. The first row attention maps are for an instance with superficial cues, and the second row attention maps are for an instance without superficial cues (Appendix E shows all layers).

Hypothesis: The doctor the president advised.
label: Entailment

If the model is sensitive to meaning, we expect the performance on this meaningless input to drop to random performance because the model was trained on meaningful input.

Figure 4a and Figure 4b shows the results of prompt-based RoBERTa trained on MNLI and SNLI, respectively. The figures show the results of a prompt-based model trained on meaning-containing instances evaluated on the test set of instances whose meaning is preserved (Yes) and when the instances are made meaningless (No). The results show that when meaning is removed from the instances, the model barely changes its predictions, indicating that the model hardly relies on the meaning of the instances.

Future Work At this point, some questions still remain unanswered: (1) What are the specific superficial cues that the models exploit? This still remains a hard interpretability question. (2) Are there any prompts that discourage models from exploiting superficial cues? (3) Can incorporating task demonstrations without superficial cues mitigate against the reliance on superficial cues?



(a) MNLI (b) SNLI

Figure 4: The average accuracy of prompt-based RoBERTa trained on MNLI (a) and SNLI (b) instances with the correct word order ("Meaningful") when evaluated on HANS dataset when the input has the correct word order (Yes in legend) and when the word order is shuffled (No in legend).

7 Related Works

Few-shot Prompting Language model prompting was popularized by recent work of GPT-3 (Brown et al., 2020). Brown et al. (2020) showed that by using prompts and some task demonstrations, GPT-3 could perform a number of tasks in the few shot setup. Following this work, Schick and Schütze (2021a) showed that even much smaller language models such as RoBERTa-large could perform well when finetuned with prompts. The subsequent work (Schick and Schütze, 2021b) demonstrated that a smaller model could achieve similar performance to GPT-3 in a few shot setup once finetuned with prompts. Many works proposed better ways of generating prompts and answer keys (Gao et al., 2021; Le Scao and Rush, 2021). Our work develops on these works to gain more insights into model predictions. Specifically, we undertake the first investigation of whether prompt-based models exploit superficial cues.

Superficial Cues in Datasets Superficial cues have been analyzed across several natural language understanding datasets. Gururangan et al. (2018) analyzed contextless superficial cues—hypothesis only superficial cues—in the MNLI dataset and SNLI dataset using fastText (Joulin et al., 2017), finding that a little over half of MNLI contain superficial cues and 67% of SNLI contain superficial cues. We argue that these figures could be outdated. Hence, we reanalyze contextless superficial cues in MNLI and SNLI. McCoy et al. (2019) analyzed context superficial cues in MNLI, finding that lexical overlap is one superficial cue that can be exploited to correctly predict entailment labels. Following their analysis, they released the HANS dataset that has an equal number of instances with superficial cues and those without superficial cues. In this work, we use the HANS dataset to evaluate the models' ability to exploit superficial cues. Kavumba et al. (2019) analyzed contextless superficial cues in COPA using RoBERTa and productivity measures introduced by Niven and Kao (2019). However, they did not analyze context superficial cues. In this work, we analyze context superficial cues in COPA, and we use the contextless superficial cues from Kavumba et al. (2019).

Exploiting Superficial Cues in Head Language Models Head language models that use a task-specific head for a downstream task have been analyzed on their ability to exploit superficial cues in datasets. McCoy et al. (2019) found that head BERT (Devlin et al., 2019) exploits superficial cues on MNLI dataset. Similarly, Niven and Kao (2019) found that head BERT exploits superficial cues on the argument reasoning comprehension task and Kavumba et al. (2019) analyzed BERT and RoBERTa's ability to exploit superficial cues on the COPA dataset. While head models have been analyzed already, prompt-based models have not been analyzed yet. In this paper, we investigated whether prompt-based models also exploit superficial cues.

8 Conclusions

We presented the first analysis of whether prompt-based models exploit superficial cues. We found that prompt-based models exploit superficial cues and fail to generalize well to instances without superficial cues on MNLI, SNLI, COPA, and HANS. We, further, proposed to divide superficial cues into two: context and contextless superficial cues. Our analysis of MNLI, SNLI, and COPA has revealed more superficial cues than was previously known.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting 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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772, New Orleans, Louisiana. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3410–3416, Hong Kong, China. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the Winograd schema challenge and SWAG. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3373–3378, Hong Kong, China. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Datasets

A.1 MNLI

The Multi-Genre Natural Language Inference (Williams et al., 2018, MNLI) dataset is an important dataset of natural language inference which is also part of the SuperGLUE benchmark (Wang et al., 2019). Given a premise, such as *"The Old One always comforted Ca'daan, except today."*, and a hypothesis, in this case, *"Ca'daan knew the Old One very well."* (neutral), a model is asked to pick one label from among three, {contradiction, neutral, entailment}. The test set of MNLI is divided into two subsets according to whether the domain of each test instance matches the domain of the training set: matched (in-domain instances) and mismatched (out-of-domain instances).

A.2 SNLI

The Stanford Natural Language Inference (Bowman et al., 2015, SNLI) is a popular natural language inference dataset with the same format as MNLI.

A.3 HANS Datasets

McCoy et al. (2019) identified three context superficial cues in MNLI: *lexical overlap*, *subsequence*, and *constituent*. These superficial cues are predictive of entailment label. Both *subsequence* and *constituent* superficial cues are special cases of lexical *lexical overlap*. Following this analysis, McCoy et al. (2019) created the HANS (Heuristic Analysis for NLI Systems) dataset, which includes instances with superficial cues—lexical overlap entails an entailment label and instances without superficial cues—lexical overlap does not necessarily mean it is an entailment label.

A.4 COPA

The Choice of Plausible Alternatives (Roemmele et al., 2011, COPA) dataset is a popular commonsense dataset that asks the model to select the most plausible answer choice from the set of two candidates, which is also a part of the SuperGLUE benchmark. For example, given a premise such as *"I tipped the bottle."* and a question that can either be the *cause* or *effect*, the model is asked to select the most plausible alternative from either *"The liquid in the bottle froze."* or *"The liquid in the bottle poured out."* (correct).

B Prompt Templates and Answer Keys

In the templates we show the part that is replaced with actual question input between two braces ({}), and we represent the mask token for a language model with <MASK>.

B.1 MNLI

We use the template and answer keys from (Gao et al., 2021).
Template:

```
{premise}? <MASK>, {hypothesis}
```

For the answer key that maps from model output to task labels we used:

```
{yes: entailment, maybe:
netural, No: contradiction}
```

B.2 SNLI

We use the template and answer keys from (Gao et al., 2021).
Template:

```
{premise}? <MASK>, {hypothesis}
```

For the *answer key* that maps from model output to task labels we used:

```
{yes: entailment, maybe:
netural, No: contradiction}
```

B.3 COPA

For COPA we use different templates from Schick and Schütze (2021b); Le Scao and Rush (2021) depending on the question type. COPA instances can have a *cause* or *effect* question.

For a *cause* question with a given *premise* and two possible causes; *choice1* and *choice2*. We use the template;

```
"{choice1}" or "{choice2}"?
   {premise}, so <MASK>.
```

and

```
{choice1} or {choice2}?
{premise}, so <MASK>.
```

For an *effect* question with a given *premise* and two possible causes; *choice1* and *choice2*. We use the template;

```
"{choice1}" or "{choice2}"?
<MASK>, because {premise}.
```

and

```
{choice1} or {choice2}?
<MASK>, because {premise}.
```

For the answer key we use the identity function.

Train set	Cue Type	W/ Superficial Cues		W/O Superficial Cues	
		Meaningful	Meaningless	Meaningful	Meaningless
MNLI	lexical_overlap	99.0 ±0.4	92.3±3.7	9.3 ±2.9	6.8 ±3.7
	subsequence	99.6 ±0.1	89.5 ±5.6	3.9 ±0.3	7.6 ±3.6
	constituent	97.4 ±1.0	89.9 ±3.8	8.9 ±2.2	12.4 ±4.7
SNLI	lexical_overlap	87.8 ±8.4	80.7 ±8.3	57.0 ±17.8	16.7 ±7.1
	subsequence	90.9 ±7.1	85.2 ±6.7	20.6 ±10.2	18.6 ±8.1
	constituent	95.3 ±3.9	75.8 ±8.6	17.6 ±15.9	25.8 ±11.3

Table 5: Average accuracy on HANS Datasets for prompt-based RoBERTa models trained on meaningful MNLI and SNLI; and evaluated on instances whose meaning has been preserved (Meaningful) and instances whose meaning has been removed (Meaningless). The instances are further divided into instances with (W/) and without (W/O) context superficial cues

Dataset	w/ s. cues	w/o s. cues	Overall
MNLI	69.5±0.1	38.6±0.6	66.7±0.1
MNLI-mm	72.0±0.2	41.5±0.7	69.1±0.2
SNLI	82.2±0.9	67.1±1.0	78.0±0.9

Table 6: Results of of prompt-based RoBERTa on MNLI and SNLI instances with contextless superficial cues in the hypothesis and instances without contextless superficial cues.

C Training Details

All our experiments are run on a single NVIDIA Tesla T4 GPU with 16GB memory. All the models are trained in a few-shot setup. Thus they training only takes a few minutes to complete. We use a learning rate of 1e-5 for all our experiments.

D Numeric Results

For all the figures containing results—figure 2 and figure 4—we also include numeric results in table 5 and table 6—to show the exact values and standard deviations to make it easier to compare with future work.

E Attention Maps

This appendix shows the complete attention maps partially visualized in section 6.2.

The attention maps are for a instance with superficial cues below.

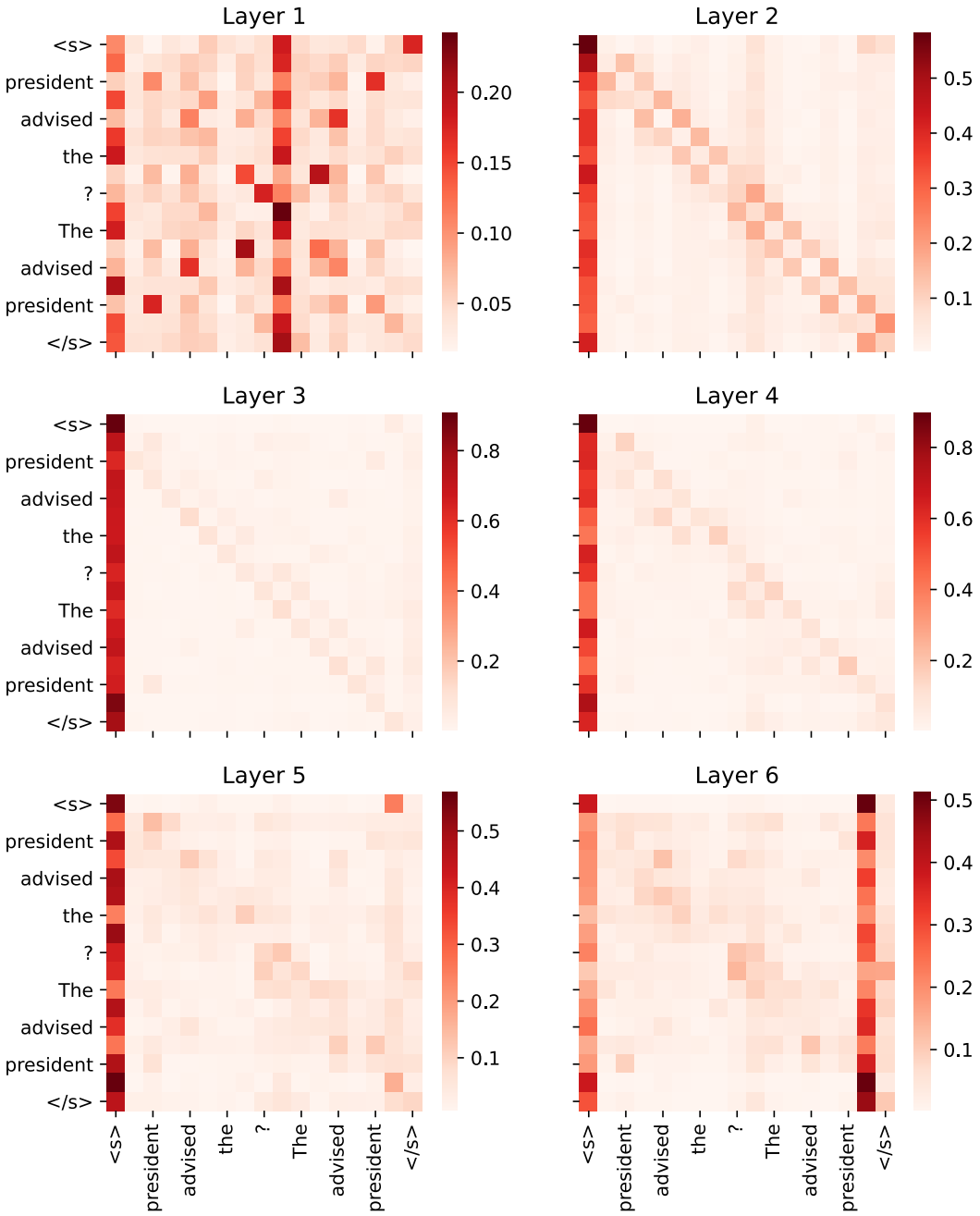
Premise: The president was advised by the doctor.
Hypothesis: The doctor advised the president.
label: Entailment

And an instance without superficial cues:

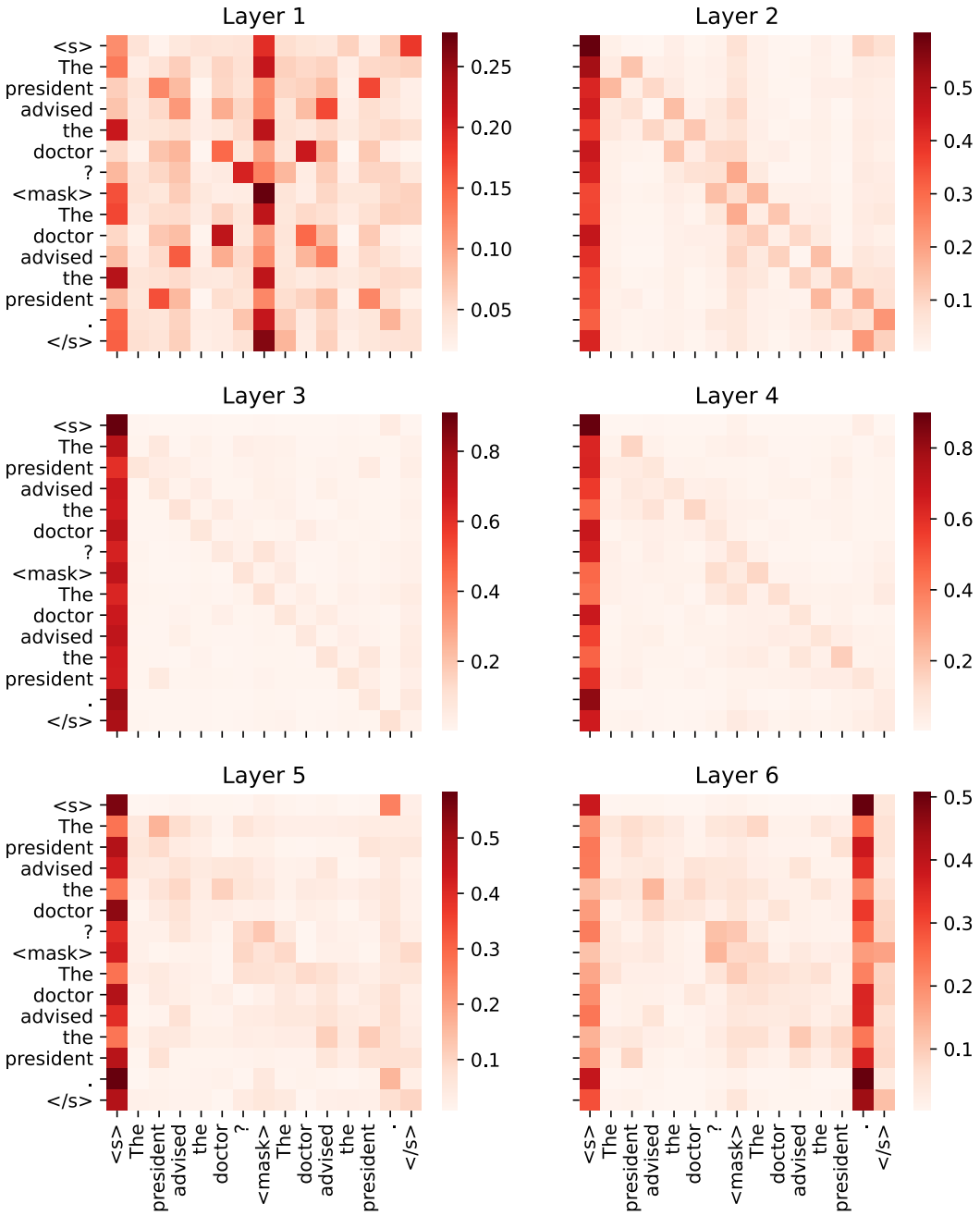
Premise: The president advised the doctor.
Hypothesis: The doctor advised the president.
label: Non-Entailment

To make it easier to compare the attention maps, we alternate the pages between the attention maps for an instance with and without superficial cues. Each page has six attention maps to make it easier to see all the points.

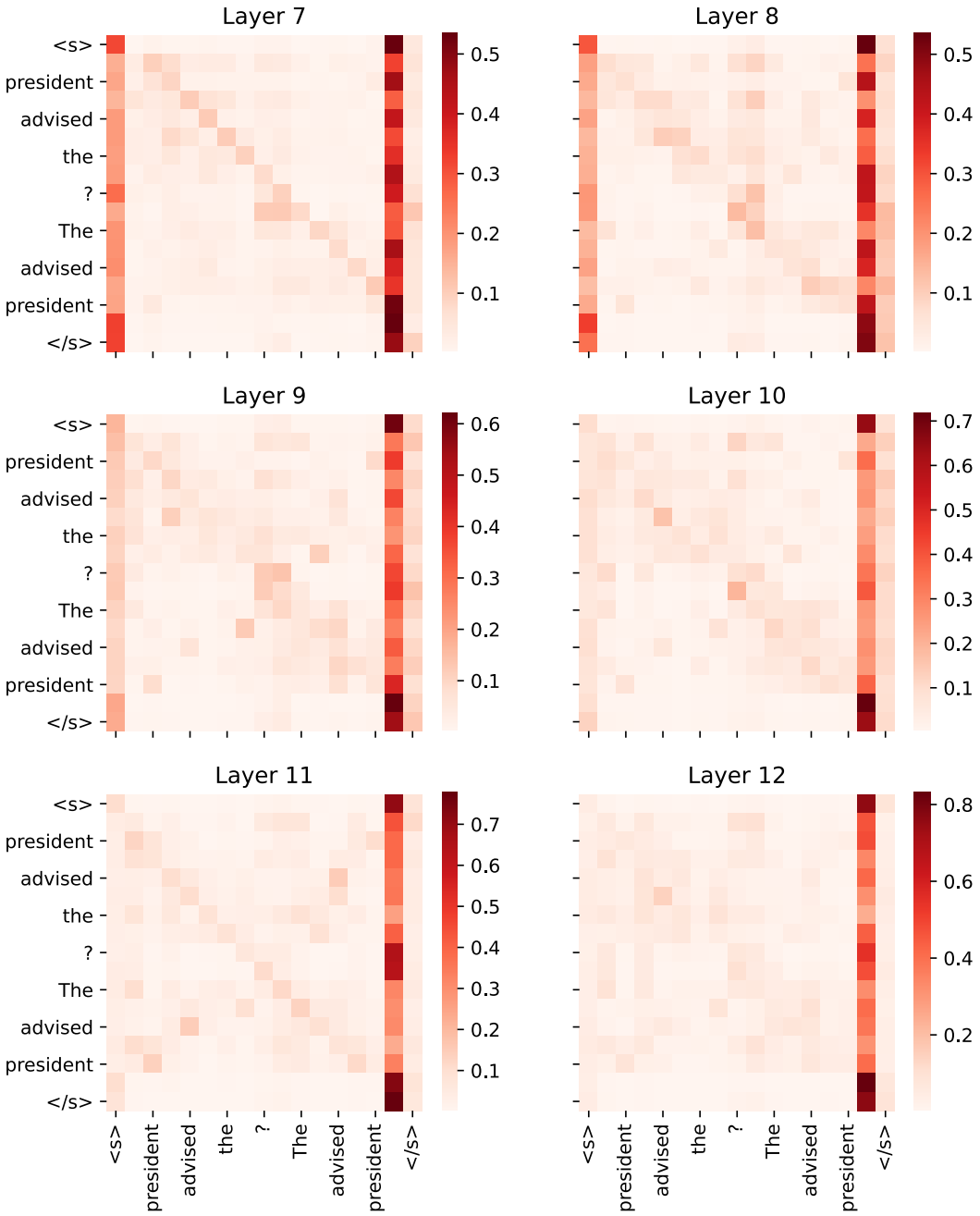
Attention Maps for NLI Instance with Superficial Cues pt. 1



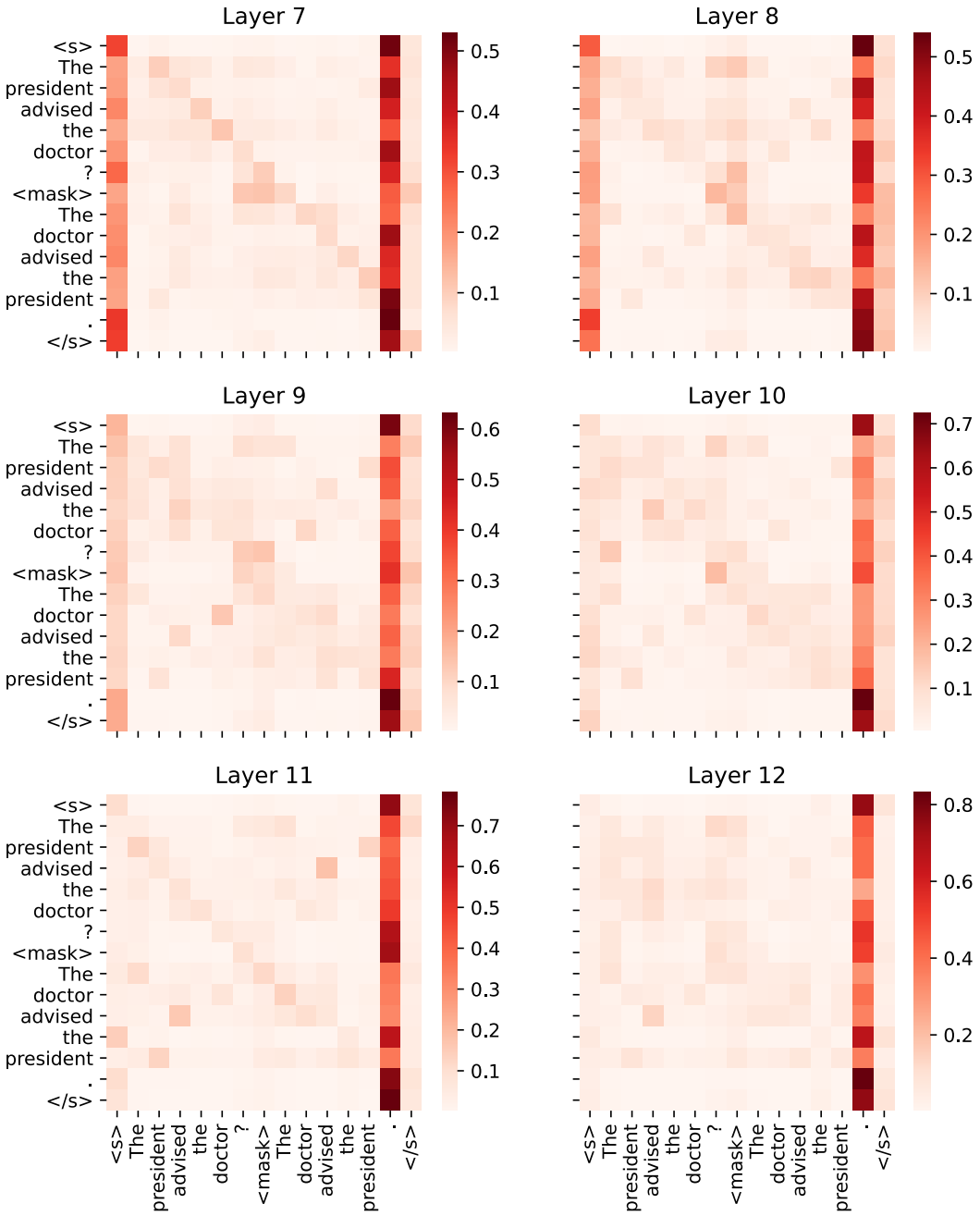
Attention Maps for NLI Instance without Superficial Cues pt. 1



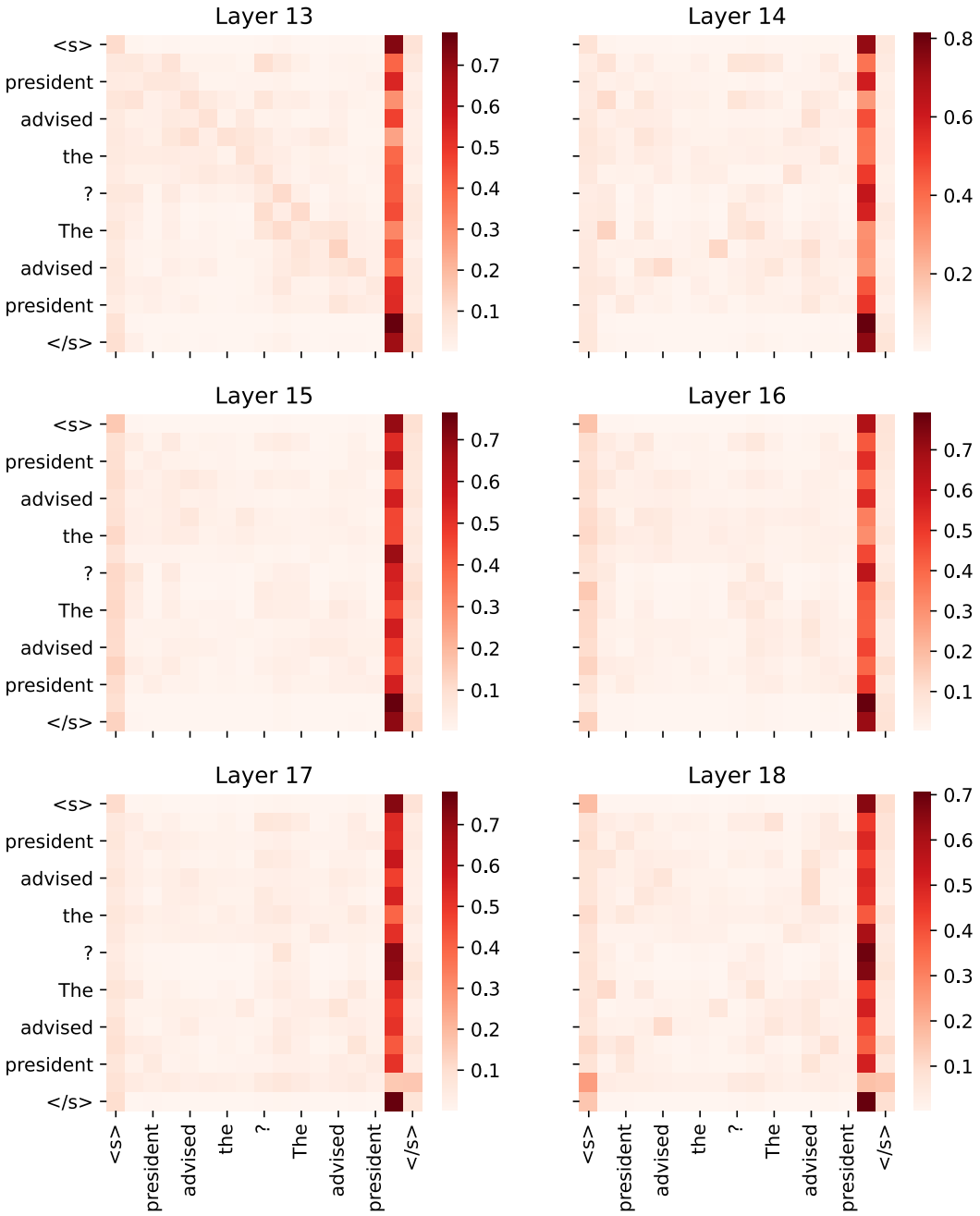
Attention Maps for NLI Instance with Superficial Cues pt. 2



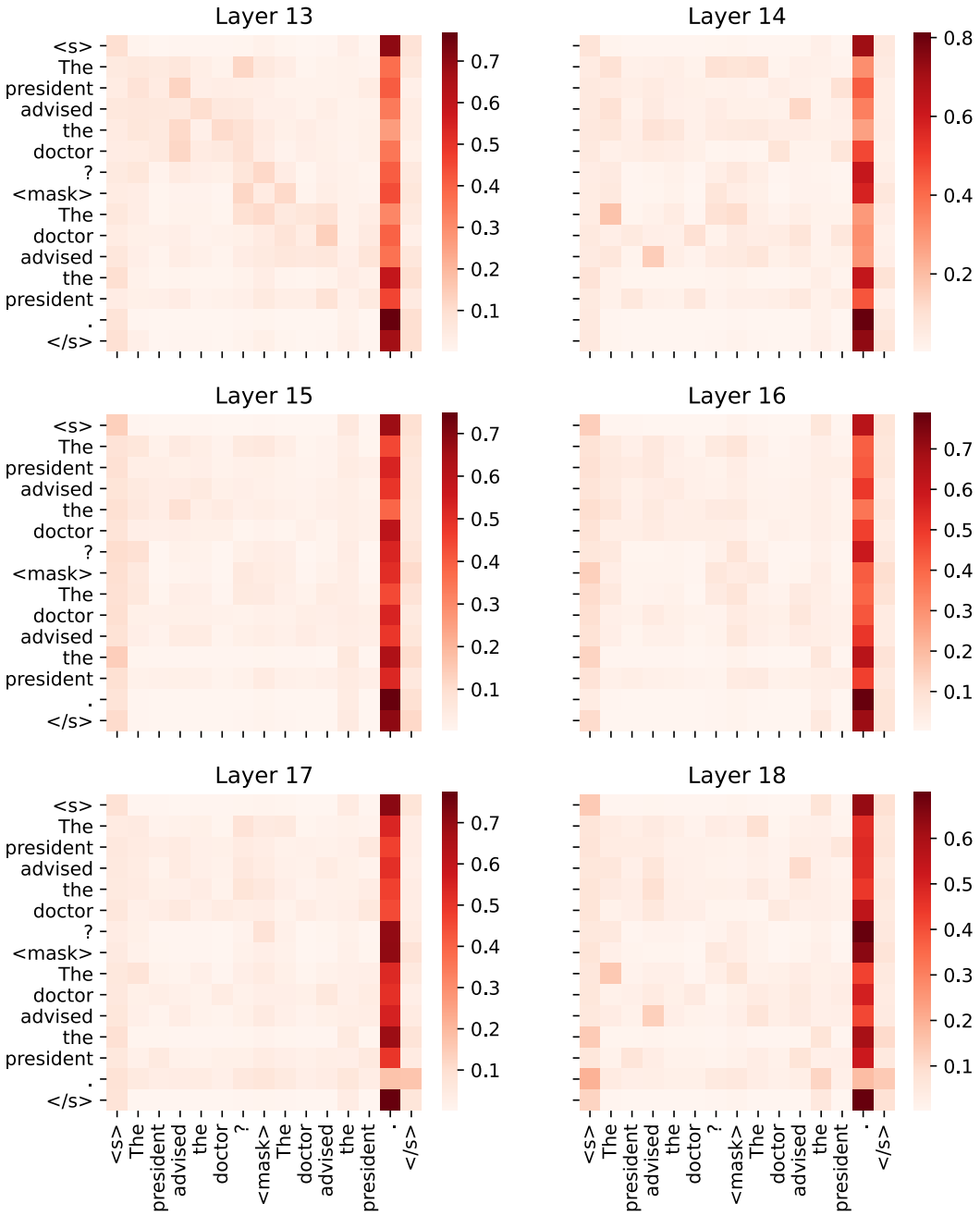
Attention Maps for NLI Instance without Superficial Cues pt. 2



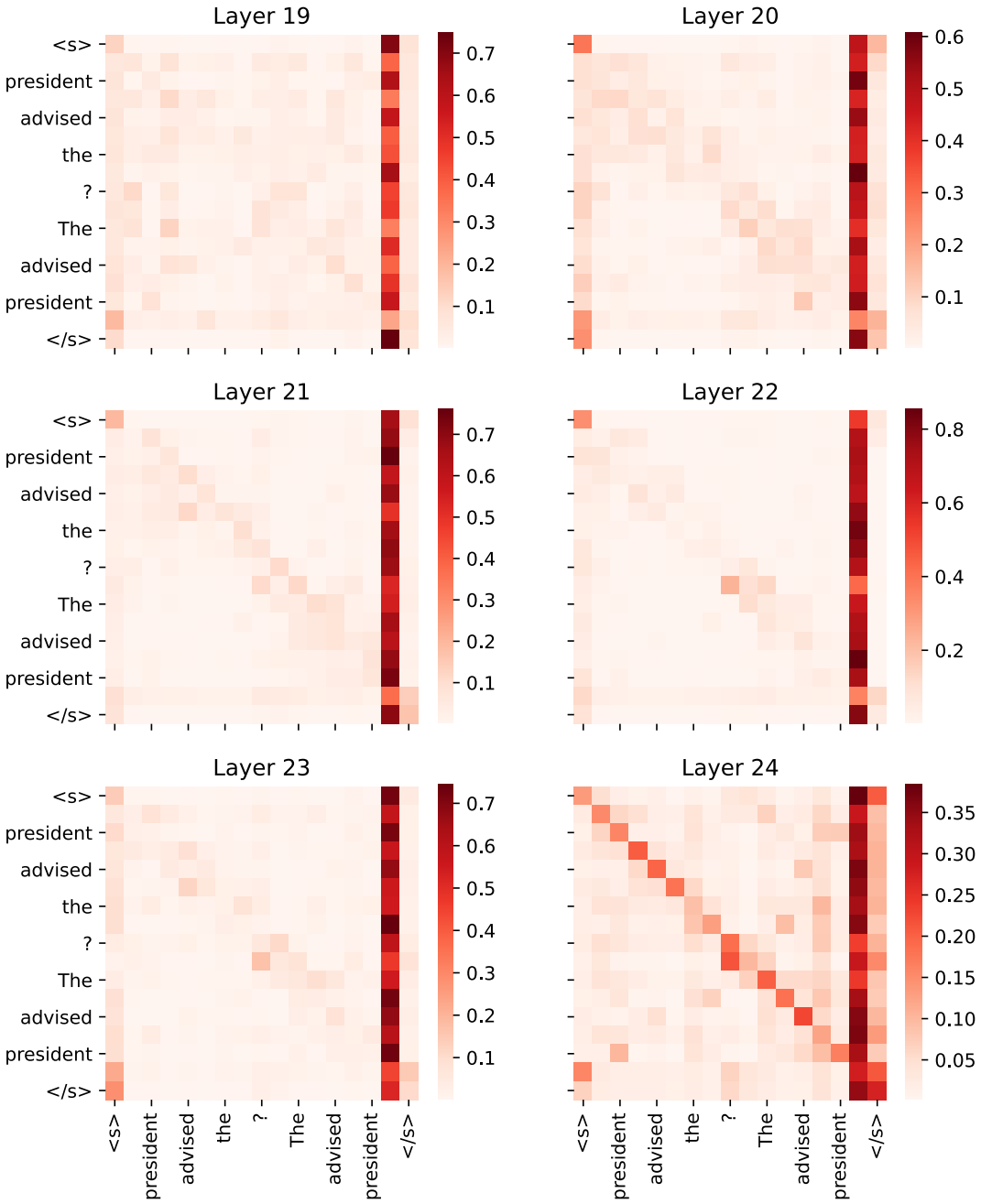
Attention Maps for NLI Instance with Superficial Cues pt. 3



Attention Maps for NLI Instance without Superficial Cues pt. 3



Attention Maps for NLI Instance with Superficial Cues pt. 4



Attention Maps for NLI Instance without Superficial Cues pt. 4