TüDuo at SemEval-2024 Task 2: Flan-T5 and Data Augmentation for Biomedical NLI

Veronika Smilga and **Hazem Alabiad**
University of Tübingen, Germany
first.last@student.uni-tuebingen.de

Abstract

This paper explores using data augmentation with smaller language models under 3 billion parameters for the SemEval-2024 Task 2 on Biomedical Natural Language Inference for Clinical Trials. We fine-tune models from the Flan-T5 family with and without using augmented data automatically generated by GPT-3.5-Turbo and find that data augmentation through techniques like synonym replacement, syntactic changes, adding random facts, and meaning reversion improves model faithfulness (ability to change predictions for semantically different inputs) and consistency (ability to give same predictions for semantic preserving changes). However, data augmentation tends to decrease performance on the original dataset distribution, as measured by F1 score. Our best system is the Flan-T5 XL model fine-tuned on the original training data combined with over 6,000 augmented examples. The system ranks in the top 10 for all three metrics[1].

1 Introduction

In the recent years, the rapid and triumphant advance of Large Language Models (LLMs) has affected virtually every area of NLP, biomedical NLP included. We aim to prove that Biomedical NLP can still benefit from smaller models of no more than three billion parameters. First, as the saying goes, "You must not use a steam hammer to crack a nut, if a nutcracker would do". In other words, while LLMs' performance is unmatched in complex applications, smaller models may be perfectly sufficient for simpler tasks, such as text classification or natural language inference. Second, being pre-trained on extremely large corpora of unlabelled data, modern LLMs have been shown to exhibit dataset-related bias (Acerbi and Stubbersfield, 2023). In fields with a high error cost, pre-training and fine-tuning models on smaller, carefully curated, high-quality datasets is safer and more predictable than using black-box giant LLMs in a zero-shot or few-shot setting. Finally, as of now, best-performing state-of-the-art LLMs are either largely unavailable to the end-user due to computational constraints (for open-source models) or cost-inefficient (for proprietary models with access via API).

The NLI4CT-2024 Shared Task (Jullien et al., 2024) consists in building a system for natural language inference (NLI) based on a collection of breast cancer Clinical Trial Reports (CTRs) in English. The task's main challenge is the complex and heterogeneous nature of the data. For each datapoint, the premise comes from one of the four sections of a CTR – Intervention, Eligibility, Results, or Adverse Events. Naturally, the sections are different from each other in terms of the mean length, the proportion of numerical data present, and the level of world knowledge required for drawing conclusions. Compared to the previous year's iteration of the task (Jullien et al., 2023), this year's challenge calls for a system robust to alterations in the data. Apart from F1 measure, two new metrics are used to evaluate the model performance: **faithfulness**, "measuring the ability of a model to correctly change its predictions when exposed to a semantic-altering intervention", and **consistency**, "measuring the ability of a system to predict the same label for original statements and contrast statements for semantic preserving interventions".

According to the last year participants' reports, various augmentation techniques have not led to significant performance improvement in terms of F1 and the top-3 best-performing systems did not use data augmentation at all (Jullien et al., 2023). However, given the new metrics that are used in this year's evaluation, it seems reasonable to continue exploring the effect that various kinds of augmentation have on F1, faithfulness, and consistency at the same time. In this paper, we fine-tune mod-

[1]Our code is available at https://github.com/smilni/semeval2024_safe_biomedical_nli

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 737–744

els of Flan-T5 family with and without the use of augmented data automatically generated using GPT-3.5-Turbo. We find that using various kinds of additional data leads to an increase in model's faithfulness and consistency, but a decrease in F1. Our best system is Flan-T5 XL, fine-tuned on 1900 original train and development instances and 6650 automatically generated ones. The system ranks 7th for consistency, 9th for faithfulness, and 10th for F1.

2 Related work

Language Models and Biomedical NLP There has been a surge of LLMs fine-tuned on biomedical data, from relatively small – 7 billion parameter ChatDoctor (Yunxiang et al., 2023), MedAlpaca (Han et al., 2023), PMC-LLAMA (Wu et al., 2023); 6 billion parameter DoctorGLM (Xiong et al., 2023) and OphGLM (Gao et al., 2023) – to extremely large ones – 540B Med-PaLM (Singhal et al., 2023), 175B Codex-Med (Liévin et al., 2022), 80B Med-Flamingo (Moor et al., 2023) – which were reported to break state-of-the-art results on a number of biomedical NLP tasks. Moreover, without any fine-tuning on biomedical data, GPT-4 was reported to have passed every step of the US-medical licensing exam (Nori et al., 2023). However, researchers argue that smaller language models, such as T5 Base and T5 Large, still outperform gigantic all-purpose models when fine-tuned for a specific task (Lehman et al., 2023).

Model Robustness in NLI tasks Many NLI models suffer from bias related to superficial correlations between input text features and labels in the training dataset, which leads to a drop in performance on datasets where these correlations do not hold (Rajaee et al., 2022). Among those are hypothesis only bias, where models rely mostly on the hypothesis without taking premise and premise-hypothesis relations into account (Poliak et al., 2018), and word-overlap bias, where models rely on the presence of shared words or phrases in premise and hypothesis (McCoy et al., 2019). Various techniques may be used to mitigate this bias, such as adversarial training (Stacey et al., 2020) and data augmentation with predicate-argument structures (Moosavi et al., 2020) and syntactic transformations (Min et al., 2020).

3 Experimental setup

In our experiments, we aim to test whether language models of relatively small size, under three billion parameters, can achieve decent performance on a task with a simple objective – as the model chooses between only two options, entailment and contradiction – and complex data – as dealing with Clinical Test Reports requires complex reasoning and understanding of numerical data. As a starting point for the experiments, we have chosen Flan-T5.

3.1 Selecting the model

Flan-T5 (Chung et al., 2022) is an updated checkpoint of T5 (Raffel et al., 2020), instruction-fine-tuned on a number of new NLP tasks, which outperforms baseline T5 models of the corresponding size on a number of benchmarks. It also features improved instruction-following capabilities and generalizes well on new tasks, not present in the training data. On the previous year's iteration of NLI4CTR, the system that featured fine-tuned Flan-T5-XXL (Kanakarajan and Sankarasubbu, 2023) without any biomedical pre-training data augmentation showed an impressive performance, ranking second.

First, we evaluate the model's performance in three scenarios – zero-shot, few-shot and after fine-tuning. Due to computational constraints we limit our experiments to language models of under three billion parameters, so we test only Flan-T5 Small (80M parameters), Flan-T5 Base (250M parameters), Flan-T5 Large (780M parameters), and Flan-T5 XL (3B parameters).

When testing the models in a zero-shot setting, we and use one of the NLI prompt templates provided by Flan-T5 developers [2]. In cases where two CTRs are given, we concatenate them using newline character as a separator. For a few-shot setting, we use the same prompt template as on the previous step, but enhance it with two hand-picked short CTR-hypothesis pairs from the training set – one with entailment and the other with contradiction relation. Refer to Appendix A for the prompts used for querying Flan-T5 in a zero-shot and two-shot setting.

Finally, we carry out fine-tuning with the use of HuggingFace Transformers library on the entire train set. The same set of hyperparameters is used for all models: `auto_find_batch_size`

[2] https://github.com/google-research/FLAN/blob/main/flan/v2/flan_templates_branched.py

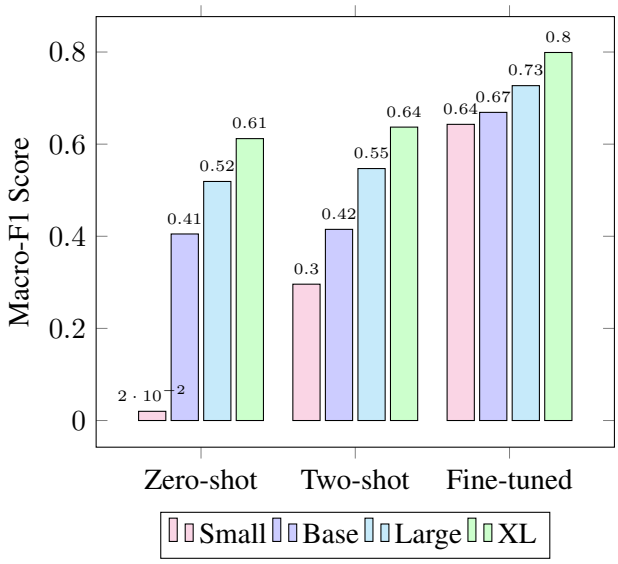

Figure 1: Flan-T5 model family performance, calculated on development dataset

`= True`, `learning_rate = 1e-3`, `optimizer = adamw_torch`. Flan-T5 Small and Flan-T5 Base are fine-tuned for 5 epochs in full precision. Due to computational constraints, Flan-T5 Large and Flan-T5 XL, however, were fine-tuned for 3 epochs using int8 precision and HuggingFace implementation of Low-Rank Adaptors (LoRA) algorithm (Hu et al., 2021).

Figure 1 summarizes the results obtained after evaluating the model on the development set in three different settings: zero-shot, few-shot, and after fine-tuning. There is a clear correlation between the model's performance and its size and between the model's performance and the number of train examples provided to it as well. In all cases, providing two examples from the training data to the model in a few-shot setting improves the performance of the model slightly, while fine-tuning it on all given training data results in a substantial performance boost. The best-performing model so far is fine-tuned Flan-T5 XL.

3.2 Data augmentation

We assume that the training data should be augmented in two key ways to create a **faithful** and **consistent** system. First, we should add paraphrased versions of the original datapoints, with semantic meaning and label preserved. It will ensure that the system is **consistent**, i.e. produces the same output for semantically equivalent inputs. Second, we should include semantically altered versions of the original datapoints, with semantic meaning changed and a reverted label assigned. It will ensure that the system is **faithful**, i.e. change its output when encountering an input semantically different to the one seen before. The presence of these three types of datapoints – original, paraphrased in a semantically preserving way, and paraphrased in a semantically altering way – is expected to improve the model's performance. We assume that these examples will teach the model to consistently handle the semantics of the sentence, mitigating the impact of superficial features like word overlap between a premise and a hypothesis on the model's performance.

We apply four types of alterations to hypotheses:

1. Synonym-based semantic-preserving changes, where certain words within a sentence are substituted with their synonymous counterparts.

2. Syntactic semantic-preserving changes, where the syntactic structure of the sentence is changed while the semantic meaning remains the same.

3. Random fact addition semantic-preserving changes, where a true random fact is appended to the hypothesis without affecting its truth value.

4. Semantic-altering changes, where a sentence contradictory to the original hypothesis is formulated.

Semantic-preserving changes 1), 2), and 3) are applied to all hypotheses, while semantic-altering change 4) is only applied to hypotheses that were

Original hypothesis	Heart-related adverse events were recorded in both the primary trial and the secondary trial. `[entailment]`
Synonym-based alteration	*Cardiovascular* adverse events were *documented* in both the primary *study* and the secondary *study*. `[entailment]`
Syntactic alteration	Both the primary trial and the secondary trial recorded adverse events related to the heart. `[entailment]`
Random fact addition	*Lymphadenopathy is the enlargement of lymph nodes due to infection, inflammation, or cancer.* Heart-related adverse events were recorded in both the primary trial and the secondary trial. `[entailment]`
Semantic-altering change	Heart-related adverse events were *not* recorded in either the primary trial or the secondary trial. `[contradiction]`

Table 1: Examples for each kind of alterations

initially labeled as entailment, changing the label to contradiction. The reason for this decision is that reverting a hypothesis that follows from some text produces a hypothesis that contradicts this text, but not vice versa. You may find examples for each kind of alterations in Table 1.

We access GPT-3.5-Turbo via OpenAI API to generate new hypotheses for each CTR-hypothesis pair, using a distinct hand-crafted prompt for each kind of alteration. You may find the text of each prompt in Appendix B. Four new hypotheses are generated for each "entailment" CTR-hypothesis pair, with both semantic-preserving and semantic-altering changes applied, and three new hypotheses are generated for each "contradiction" CTR-hypothesis pair, with only semantic-preserving changes applied. In all cases, CTR text itself remains unaltered, and only hypothesis is affected.

As a result, we obtain 3400 new entries for 850 original train CTR-hypothesis pairs labelled as entailment and 2550 new entries for 850 original train CTR-hypothesis pairs labelled as contradiction. The process of generating 5950 data points, thus increasing our dataset by 4.5 times, cost $0.86 and took 1.5 hours to complete.

4 Results

4.1 Individual Augmentation Analysis

We fine-tune Flan-T5 XL model on augmented data using the same set of hyperparameters as in Section 3.1. First, we fine-tune the model separately on each type of augmented data (combined with the original data) to estimate how augmentation of each kind affects the performance. The results are presented in Table 2.

Interestingly, only one kind of augmentation, the synonym-based one, had a positive effect on the model's performance on the original dataset, while the others led to a decrease in F1. All kinds of augmentations resulted in a model with higher consistency, i.e. a model better at producing the same output for hypotheses with the same meaning. The alteration that consisted in adding random true facts to hypotheses led to the highest increase in consistency. However, only semantic-altering change resulted in a more faithful model, i.e. a model better at changing its prediction when encountering a similar but semantically different hypothesis. All semantic-preserving changes led to a decrease in the model's faithfulness. Overall, our data augmentation techniques have proven to be efficient in improving the model's robustness. However, they have simultaneously resulted in a worse performance on the original data.

4.2 Final Model Selection

The next step was to try out different combinations of augmented data to reach the optimal performance in terms of the largest increase in both faithfulness and consistency and the smallest decrease in terms of F1. As the goal of the competition was to create a faithful and consistent system, we prioritized these metrics over F1 when choosing the model for the final submission. Thus, we chose the model trained on the entire set of augmented data that demonstrates higher faithfulness and consistency but lower F1. For the final submission, we additionally enriched the dataset with 200 more entries from development data and 700 new augmented entries created using techniques described in Section 3.2. The results obtained after fine-tuning the model on the entire augmented dataset are presented in Table 3.

	F1	Faithfulness	Consistency
Original train data only	0.779	0.780	0.667
Original train data + synonym-based alterations	**0.780**	0.715	0.681
Original train data + syntax-based alterations	0.764	0.748	0.698
Original train data + random facts addition	0.748	0.736	**0.725**
Original train data + reverted meaning alterations	0.735	**0.854**	0.686

Table 2: Flan-T5 XL performance when trained on different kinds of augmented data, calculated on test dataset

	F1	Faithfulness	Consistency
Original train data only	**0.779**	0.780	0.667
Original train data + all augmented train data	0.745	**0.851**	0.748
Original train and dev data + all augmented train and dev data	0.760	0.841	**0.752**

Table 3: Flan-T5 XL performance when trained on all kinds of augmented data, calculated on test dataset

4.3 Other approaches

Numerical inference is a known challenge for large language models. We assumed that the model's performance might vary across different CTR sections, with a decrease in performance for sections that contain most numbers. To check this assumption, we calculated the final model's F1 for each section separately. Calculations were performed on the development dataset as we had no access to test dataset labels during the development and evaluation stages. The results are presented in Table 4.

	F1
Adverse Events	0.711
Eligibility	0.821
Intervention	0.861
Results	0.759
All sections	0.783

Table 4: Final model's performance on each CTR section, calculated on development dataset

Adverse events, the section that, according to our observations, most often contained numbers in premise as well as hypothesis and required numerical inference to determine the relation between them, had the lowest F1 of all.

We attempted to develop a separate model, Flan-T5 XL with the same hyperparameter set as in Section 3.1, to tackle CTR-hypothesis pairs of this kind. The model was first pre-fine-tuned on EQUATE dataset (Ravichander et al., 2019) for 3 epochs in an attempt to enhance its numerical inference capabilities. Then it was further fine-tuned on the original and augmented CTR-hypothesis pairs of Adverse Events category for 3 epochs as well. We then used the original model to produce predictions for Eligibility, Intervention, and Results sections and the new model to produce predictions for Adverse Events section. However, on test data, this approach resulted in a decrease in performance with an F1 of 0.756 (-0.004), faithfulness of 0.781 (-0.06) and consistency of 0.722 (-0.031). We suppose that the decrease in performance is explained by the fact that the second model, trained on ~1/4 of all data (only one section out of four), simply did not encounter enough data to develop robustness comparable to that of the final model trained on the entire dataset.

5 Conclusion

In this paper, we explore the impact of data augmentation on model performance and robustness. Specifically, we focus on leveraging advanced language models like GPT-3.5-Turbo to expand the training set for fine-tuning smaller models such as Flan-T5 XL. Our experiments involve various prompts to generate new CTR-hypothesis pairs. Enriching the training set with new examples that underwent semantic-preserving changes, such as synonym replacement, change in word or clause

order, and random true fact addition, improves the model's consistency. Adding augmented examples that underwent semantic-altering changes, such as meaning reversion, improves the model's faithfulness as well as consistency. However, all kinds of augmentation except for synonym replacement lead to a decrease in model performance in terms of F1 on the original unaltered dataset. The model selected for the final submission is Flan-T5 XL fine-tuned on augmented development and training set. It features higher robustness but lower base performance than Flan-T5 XL fine-tuned on original data only, with faithfulness of 0.841 (+0.061), consistency of 0.752 (+0.085), and F1 of 0.76 (-0.019).

Acknowledgements

We would like to thank Dr. Çağrı Çöltekin for guiding, supporting and inspiring us on our way.

References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. 2023. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. *arXiv preprint arXiv:2305.02993*.

Kamal Raj Kanakarajan and Malaikannan Sankarasubbu. 2023. Saama ai research at semeval-2023 task 7: Exploring the capabilities of flan-t5 for multi-evidence natural language inference in clinical trial data. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 995–1003.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Nafise Sadat Moosavi, Marcel de Boer, Prasetya Ajie Utama, and Iryna Gurevych. 2020. Improving robustness by augmenting training sentences with predicate-argument structures. *arXiv preprint arXiv:2010.12510*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. *arXiv preprint arXiv:2211.03862*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. *arXiv preprint arXiv:2004.07790*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

A Prompts used in querying Flan-T5

```
Read the text and determine if the sentence
is true:
{premise}
Sentence: {hypothesis}
["yes", "no"]
```

Figure 2: Zero-shot prompt used in querying Flan-T5.

```
Read the text and determine if the sentence
is true:
Inclusion Criteria: Estrogen receptor or
progesterone receptor positive breast cancer
Premenopausal with regular menstrual cycles
Exclusion Criteria: Current oral
contraceptives
Sentence:  Males are not eligible for the
primary trial.
["yes", "no"]
Answer: yes

(another instruction, CTR, and answer - this
time with a contradiction relation)

Read the text and determine if the sentence
is true:
{premise}
Sentence: {hypothesis}
["yes", "no"]
Answer:
```

Figure 3: Two-shot prompt used in querying Flan-T5.

B Prompts used to obtain augmented data from GPT-3.5-Turbo

```
TEXT: {text}
Paraphrase TEXT using synonyms while
preserving its original meaning. Always keep
words "primary trial" and "secondary trial"
if present in TEXT, do not replace "primary
trial" and "secondary trial" with synonyms.
Return only paraphrased text and nothing
else.
```

Figure 4: Prompt used to generate a synonym-based paraphrased version of hypothesis.

```
TEXT: {text}
Change the syntactic structure of TEXT while
preserving its original meaning. Always keep
words "primary trial" and "secondary trial"
if present in TEXT, do not replace "primary
trial" and "secondary trial" with synonyms.
Return only paraphrased text and nothing
else.
```

Figure 5: Prompt used to generate a syntax-based paraphrased version of hypothesis.

```
Generate a random short true definition of a
random medical term. Use format '{term} is
{definition}'. Definition must be no longer
than one sentence."
```

Figure 6: Prompt used to generate a random biomedical fact to then append to hypothesis.

```
TEXT: {text}
Revert the original meaning of TEXT. The
result must contradict TEXT. Return only the
result and nothing else.
```

Figure 7: Prompt used to generate sentence with meaning contradicting that of hypothesis.