Saama Technologies at SemEval-2024 Task 2: Three-module System for NLI4CT Enhanced by LLM-generated Intermediate Labels

Hwanmun Kim, Kamal Raj Kanakarajan, Malaikannan Sankarasubbu
Saama Technologies
{hwan.kim, kamal.raj, malaikannan.sankarasubbu}@saama.com

Abstract

Participating in SemEval 2024 Task 2, we built a three-module system to predict entailment labels for NLI4CT, which consists of a sequence of the query generation module, the query answering module, and the aggregation module. We fine-tuned or prompted each module with the intermediate labels we generated with LLMs, and we optimized the combinations of different modules through experiments. Our system is ranked 19th∼24th in the SemEval 2024 Task 2 leaderboard in different metrics. We made several interesting observations regarding the correlation between different metrics and the sensitivity of our system on the aggregation module. We performed the error analysis on our system which can potentially help to improve our system further.

1 Introduction

While the surge of Large Language Models (LLMs) revolutionized many NLP tasks including Natural Language Inference (NLI), the threat of hallucination of LLMs is becoming a critical obstacle to applying these models to real-world problems. This is particularly the case for the field where the entailment of the statements with the records is crucial, such as biomedical applications dealing with clinical trials. SemEval 2024 Task 2 (Jullien et al., 2024) exactly tackles this question by providing the challenge for the NLI4CT dataset (Jullien et al., 2023a) composed of evidence made out of clinical trials, statements made based on this evidence, and the entailment labels of these statements. NLI4CT deals with the language of English only.

This paper explains our approach to the SemEval 2024 Task 2 in detail. We notice that many of the statements in the NLI4CT dataset require a series of questions to be answered before verifying the agreement of the statements with the evidence. For that, we built the three-module system which contains the query generation module, the query answering module, and the aggregation module. For the development of each module, we needed intermediate labels about queries and their answers, and we sourced these labels by few-shot and zero-shot prompting on the LLMs. Once we sourced the intermediate labels, we made a series of few-shot prompts with these intermediate labels to use them to prompt on fine-tuned models or the plain pre-trained models. Then we selected our best-performing model by experimenting with different combinations of modules.

As the result of our experiments, we built the system ranked 19th∼24th in the SemEval 2024 Task 2 leaderboard in different metrics. Despite relatively lower performance, we made several interesting observations. First, our experiments indicate that semantic-preserving interventions are more difficult to detect than semantic-altering interventions. Second, we deduce that semantic-altering intervention from negative entailment to positive entailment is more difficult to capture than the opposite, based on our experiment results. We also performed some error analysis on our system and obtained some interesting insights to improve our system in the future.

2 Background

2.1 Task description

SemEval 2024 Task 2 aims to develop a binary classification system that determines if a statement about a Clinical Trial Report (CTR) agrees with the given CTR. This challenge works on the NLI4CT dataset which contains 999 CTRs in total. Each CTR belongs to one of 4 sections (eligibility, intervention, adverse events, and results) and 2 types (single and comparison). Each CTR deals with up to 2 cohorts and includes data about up to 2 trials (1 for single and 2 for comparison). NLI4CT dataset is composed of a train set (1700 statements) and a dev set (200 statements). The test set of this year's

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1435–1457
June 20-21, 2024 ©2024 Association for Computational Linguistics

competition contains 5500 statements in total, including contrast statements to compute faithfulness and consistency (see Section 4.3).

2.2 Related Works

The rise of LLMs impacted a lot on NLI (Huang and Chang, 2023; Zhu et al., 2024; Zhao et al., 2023; Lee et al., 2023; Qin et al., 2023) and it can be particularly effective for the NLI tasks requiring the understanding of long contexts (Sanyal et al., 2024). Yet, NLI with LLMs is still struggling with issues like hallucination (Ye et al., 2023; Chen and Shu, 2023), shortcut learning (Geirhos et al., 2020; Du et al., 2023), and factual inconsistency (Fierro and Søgaard, 2022), and therefore there have been numerous efforts to overcome these drawbacks (Guu et al., 2020; Shen et al., 2023; Li et al., 2023; Varshney et al., 2023; Manakul et al., 2023; Ye et al., 2023).

While these approaches used a wide variety of architectures, the complicated natures of NLI tasks often forced many systems to adopt multiple modules designed to tackle different subtasks. In Shen et al. (2023), entailment explanations generated on a module are passed to another module to better predict the entailment between the news articles and their headlines. Guu et al. (2020) used a neural knowledge retriever that retrieves relevant pieces of information out of the corpus to assist the question-answering process in another module. Lei et al. (2023) proposed to use separate hallucination detection modules at the sentence level and at the entity level to better recognize the hallucination of the given text.

Multi-module system has advantages that each subtask is easier to optimize than the entire task as a whole. With this inspiration, we built a sequence of modules to process different subtasks in SemEval 2024 Task 2.

3 System overview

Our system is composed of three modules: the Query Generation (QG) module, the Query Answering (QA) module, and the Aggregation (AG) module (Figure 1). The query generation module generates a set of queries required to verify if the given statement is true or false. Each generated query is passed to the query answering module along with the given evidence and generates an answer for that query. Then all of the query-answer pairs are appended together with the given evidence and passed to the aggregation module to produce the entailment label prediction.

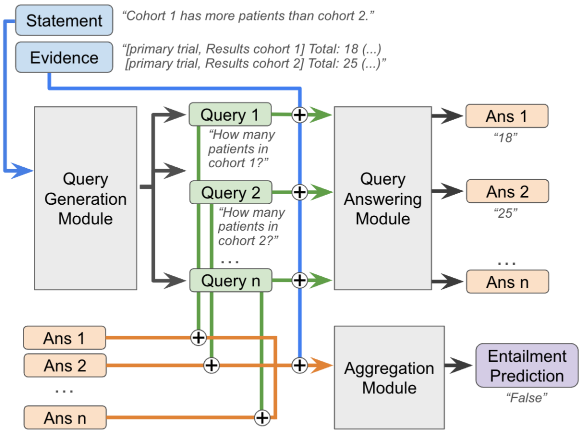
Statement
"Cohort 1 has more patients than cohort 2."
Evidence
"[primary trial, Results cohort 1] Total: 18 (...)
[primary trial, Results cohort 2] Total: 25 (...)"
Query Generation Module
Query 1
"How many patients in cohort 1?"
Query 2
"How many patients in cohort 2?"
...
Query n
Query Answering Module
Ans 1
"18"
Ans 2
"25"
...
Ans n
Ans 1
Ans 2
...
Ans n
Aggregation Module
Entailment Prediction
"False"

Figure 1: Overview of our three-module system. A sample case is illustrated with italic texts.

3.1 Intermediate label generation

To improve each module beyond the zero-shot prompting, one needs to either train/fine-tune the model or few-shot prompt with in-context examples sampled from the dataset. In either case, one needs a dataset of query-answer pairs per each statement in the train set of NLI4CT. For this, we ran our three-module system with a hand-crafted few-shot prompt for the query generation (Prompt 1) and zero-shot prompts for the other two modules (Prompt 2 and 3), on the pre-trained LLMs. Then we collected the generated byproduct query-answer pairs. To maintain the quality of labels reasonably, we only collected intermediate labels from the runs whose final entailment prediction agrees with the given entailment label in the train set. Through some experimentation, we settled down to the 987 statements with intermediate labels, which are produced by `gemini-pro` (Team et al., 2023) and `mistral-7B-instruct-v0.2`[1] (Jiang et al., 2023).

```
### For each given statement, generate queries
    to verify if the given statement is true.

You said: The majority of patients in the
    primary trial did not experience Left
    ventricular dysfunction.
To verify it,
I checked: What is the percentage of patients
    experienced Left ventricular dysfunction?
I checked: What is the percentage of patients
    had any adverse event?

(7 more such examples)
```

---

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

```
You said: (given statement)
To verify it,
```

Prompt 1: 8-shot prompt for query generation. The full list of examples is available in Appendix C.

```
### For a given question, answer the question
    based on the given evidence.
### If the question cannot be answered, respond
    as "Not Available".

#########################################
*** Question: (given question)
*** Evidence:
(given evidence)
#########################################
From this evidence, answer to the question is:
```

Prompt 2: Zero-shot prompt for query answering.

```
### For a given statement and multiple question-
    answer pairs, verify if the statement agrees
     with the question-answer pairs.
### Respond as "Agreement: True" if the
    statement agrees with the question-answer
    pairs, and respond as "Agreement: False"
    otherwise.

#########################################
*** Statement: (given statement)
*** Question1: (question 1)
*** Answer1: (answer to the question 1)
(further question-answer pairs, if exists)
#########################################
Agreement:
```

Prompt 3: Zero-shot prompt for aggregation.

3.2 Experimented modules

To figure out the best combination of modules for our system, we experimented following modules:

- **pre-trained (QG, QA, AG)**: `mistral-T5-7B-v1`[2] model with no fine-tuning. Prompted with Prompt 1 (for QG), Prompt 2 (for QA), or Prompt 3 (for AG).
- **fine-tuned (QG, QA, AG)**: `mistral-7B-instruct-v0.2` model fine-tuned with the generated intermediate labels. Prompt 1 (for QG), Prompt 2 (for QA), or Prompt 3 (for AG) is used in both training and inference processes.
- **fine-tuned, custom (QG)**: `mistral-7B-instruct-v0.2` model fine-tuned with the generated intermediate labels. Instead of global examples in Prompt 1, 8 different sets of examples customized for each section ID (eligibility, intervention, adverse events, results) and each type (single, comparison) are used according to the section ID and type of the input data, for both training and inference processes. These 8 sets of examples are sampled and modified from the generated intermediate labels and are listed in Appendix C. Instead of fine-tuning a single model for the query generation, we fine-tuned 4 different models for each section ID.
- **pre-trained, custom (QA, AG)**: `mistral-T5-7B-v1` model with no fine-tuning. Instead of the zero-shot prompts, 4 different few-shot prompts (such as Prompt 4 and 5) with examples respectively sampled and modified for each section ID from the generated intermediate labels are used. The entire examples are available in Appendix C.

```
### For a given question, answer the question
    based on the given evidence.
### If the question cannot be answered, respond
    as "Not Available".

#########################################
*** Question: (example 1: question)
*** Evidence:
(example 1: evidence)
*** Answer:
Let's think step by step. From evidence "(
    quotation from above evidence)", (reasoning
    to the answer).
Therefore, answer is "(example1: answer)"
#########################################
(more examples like above)
#########################################
*** Question: (given question)
*** Evidence:
(given evidence)
*** Answer:
Let's think step by step.
```

Prompt 4: Few-shot prompt for query answering, for the custom fine-tuned module.

```
### For a given statement and multiple question-
    answer pairs, verify if the statement agrees
     with the question-answer pairs.
### Respond as "Agreement: True" if the
    statement agrees with the question-answer
    pairs, and respond as "Agreement: False"
    otherwise.

#########################################
*** Statement: (example 1: statement)

[Pair 1]
*** Question: (example 1: question 1)
*** Answer: (example 1: answer 1)

(more question-answer pairs, if exists)
```

---

[2] https://huggingface.co/ignos/Mistral-T5-7B-v1

```
*** Response: Let’s think step by step.
(reasoning to the answer).
Therefore, the statement is (true or false).
Agreement: (true or false)

#########################################
(more examples like above)
#########################################
*** Statement: (given statement)

[Pair 1]
*** Question: (given question 1)
*** Answer: (answer to the given question 1)

(more question-answer pairs, if exists)

*** Response:
Let’s think step by step.
```

Prompt 5: Few-shot prompt for aggregation, for the custom fine-tuned module.

4 Experimental setup

4.1 Prompting for the intermediate label generation

For the intermediate label generation, we used several LLMs. For the query generation and the aggregation, we used `gemini-pro` (Team et al., 2023) serviced through Google AI Python SDK[3] with temperature 0.5. For the query answering, we used `mistral-7B-instruct-v0.2` through Huggingface Transformers[4] with temperature 0 and maximum output token 512. Further details of data processing for prompting is available in Appendix A.

4.2 Fine-tuning experiments

For resource-efficient experimentation, we adopted low-rank adaptation (LoRA) (Hu et al., 2021) for our fine-tuning experiments. We use AdamW optimizer (Loshchilov and Hutter, 2017) to optimize cross-entropy loss with label smoothing (Pereyra et al., 2017). For the experiments, we used libraries from Huggingface (Transformers, PEFT[5], TRL[6]) to deploy and fine-tune open-sourced models. Hyperparameters for the experiments are available in Appendix B. All models are fine-tuned using a 4× Quadra RTX 8000 (48GB VRAM) card.

4.3 Evaluation metric

In SemEval 2024 Task 2, the following evaluation metrics are used (Jullien et al., 2024):

- **F1-score**[7] on the binary entailment label

- **Faithfulness**: For given N statements x_i in the contrast set (C), their respective original statements y_i, and the prediction model f, faithfulness is computed as $N^{-1}\sum_1^N |f(y_i) - f(x_i)|$ where $x_i \in C : \text{Label}(x_i) \neq \text{Label}(y_i)$, and $f(y_i) = \text{Label}(y_i)$.

- **Consistency**: For the same $\{x_i\}$, $\{y_i\}$, and f above, consistency is computed as $N^{-1}\sum_1^N [1 - |f(y_i) - f(x_i)|]$ where $x_i \in C : \text{Label}(x_i) = \text{Label}(y_i)$.

5 Results

5.1 Experiments over different modules

We present the experiment results over different modules on the test set in Table 1. The entry we submitted for the SemEval 2024 Task 2 leaderboard was experiment 8, which was ranked the 23rd in F1, the 24th in consistency, and the 19th in faithfulness. Yet, we achieved our best scores in each metric in different experiments: we achieved our best F1 from experiment 3 (would be ranked 23rd if submitted) and also achieved our best consistency and faithfulness from experiment 2 (would be ranked 20th/16th if submitted). Overall, the performance of our system cannot surpass many other systems from the previous year’s competition (Jullien et al., 2023b).

By examining the results in Table 1, we can make following observations:

- For most combinations, faithfulness is higher than consistency. This may indicate that semantic-preserving interventions are more difficult to detect than semantic-altering interventions.

- F1 score and faithfulness are negatively correlated (correlation coefficient: -0.57). This comes from the negative correlation between faithfulness and precision score (correlation coefficient: -0.71). This means that the faithfulness would increase as the overall ratio of positive prediction increases, which may indicate that semantic-altering intervention from negative entailment to positive entailment is more difficult to capture than the opposite.

---

[3] https://github.com/google/generative-ai-python.

[4] https://github.com/huggingface/transformers

[5] https://github.com/huggingface/peft

[6] https://github.com/huggingface/trl

[7] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Exp #	QG Module	QA Module	AG Module	F1	Consistency	Faithfulness
1	fine-tuned	fine-tuned	fine-tuned	0.634	0.545	0.479
2	pre-trained	pre-trained	pre-trained	0.603	**0.657**	**0.781**
3	fine-tuned, custom	pre-trained	pre-trained	**0.666**	0.569	0.544
4	fine-tuned, custom	pre-trained	fine-tuned	0.525	0.589	0.748
5	fine-tuned, custom	fine-tuned	pre-trained	0.631	0.555	0.581
6	fine-tuned, custom	fine-tuned	fine-tuned	0.424	0.575	0.763
7	fine-tuned, custom	pre-trained	pre-trained, custom	0.565	0.629	0.751
8	fine-tuned, custom	pre-trained, custom	pre-trained	0.659	0.579	0.595
9	fine-tuned, custom	pre-trained, custom	pre-trained, custom	0.614	0.652	0.758

Table 1: Results of experiments over different QG, QA, and AG modules.

- Switching AG module gives the most consistent changes in the performance. As switching from pre-trained to custom pre-trained to fine-tuned, faithfulness increases and F1 decreases. Such a tendency is not so evident in QA modules.

5.2 Error analysis

To better understand what our system is lacking, we analyzed some responses from our modules used for experiment 8 from Table 1. The following are examples we find particularly interesting.

```
*** Statement: in the primary clinical trial,
    97% of patients reported experiencing
    adverse events.

[Pair 1]
*** Question: What is the percentage of patients
     in the primary clinical trial that reported
     experiencing adverse events?
*** Answer: 98%

*** Response: Let's think step by step.
(The following is the response from AG module)
The statement says that 97% of patients reported
     experiencing adverse events.
Pair 1 asks for the percentage of patients that
    reported experiencing adverse events.
The answer in Pair 1 is 98%, which is slightly
    different from the statement.
However, both the statement and the answer are
    close in terms of the percentage.
So, we can say that the statement agrees with
    the question-answer pair to some extent.
Agreement: Partially True
```

Like the above response, modules often fail to make correct inferences related to numerical results although enough information is provided. Since our approach is a sequence of modules, it might be beneficial to introduce some rule-based systems or LLMs trained/prompted to generate code to handle the numerical inputs.

```
*** Question: Does the primary clinical trial
    administer docetaxel to their patient
    cohorts?
*** Evidence:
[primary trial, INTERVENTION cohort 1]
    Chemotherapy With GM-CSF
[primary trial, INTERVENTION cohort 1]
    Doxorubicin and Cyclophosphamide (AC) with
    Granulocyte-macrophage colony-stimulating
    factor (GM-CSF) (days 4-13) Followed by
    Weekly Carboplatin/Nab- Paclitaxel
[secondary trial, INTERVENTION cohort 1]
    Neoadjuvant Chemotherapy
[secondary trial, INTERVENTION cohort 1]
    Doxorubicin (Adriamycin) + cyclophosphamide
    (Cytoxan) with pegfilgrastim or filgrastim
    growth factor support every 2 weeks for 4
    cycles, followed by docetaxel + lapatinib
    for four 21-day cycles, followed by surgery.
     Dexamethasone was administered twice-a-day
    for 3 days, starting 24 hours before the
    docetaxel infusions. After surgery +/-
    radiation, participants may receive
    trastuzumab (Herceptin) for a year.
*** Answer:
Let's think step by step.
(The following is the response from QA module)
From evidence "[secondary trial, INTERVENTION
    cohort 1] Neoadjuvant Chemotherapy",
the trial administers docetaxel + lapatinib for
    four 21-day cycles.
Therefore, answer is "Yes"
```

As shown in the above response, our QA module often hallucinates, and it is more prone to do so when the relevant information is missing in the evidence about the current trial/cohort while the relevant information exists in the other trial/cohort.

6 Limitations

Our investigation of fine-tuned models is restricted to relatively small open-sourced models due to the limited resources. The smaller size of our training data from our LLM-generated intermediate labels compared to the original train set could impact the performance of modules as well. For prompting for QA and AG modules, the number of examples is limited to 3 due to the restricted context size.

7 Conclusion

We built the three-module system to predict entailment labels for NLI4CT. While our system's performance is not superb, we made several interesting observations regarding the correlation between different metrics and the sensitivity of our system on the aggregation module. As per our error analysis, it would be interesting to introduce some rule-based systems or LLMs trained/prompted to generate code to handle the numerical inputs for the aggregation module. It would be also promising to filter the evidence input by cohorts and trials related to the question for the QA module. It would be also interesting to extend our approach to a greater variety of larger LLMs or to expand the number of examples using LLMs with larger context sizes.

References

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Minds versus machines: Rethinking entailment verification with language models.

Jiaming Shen, Jialu Liu, Dan Finnie, Negar Rahmati, Mike Bendersky, and Marc Najork. 2023. “why is this misleading?”: Detecting news headline hallucinations with explanations. In *Proceedings of the ACM Web Conference 2023*, WWW ’23, page 1662–1672, New York, NY, USA. Association for Computing Machinery.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey.

A Details of data processing

Responses from all modules are post-processed to extract the queries/answers/entailment predictions. Since the LLMs tend to repeat and continue the similar type of examples demonstrated in the prompt, we truncate all the responses up to the part where the ’#’ appears first in each response. Further post-processing was done for each module:

- **QG module**: Split the truncated response into each line, and only collected texts followed by ’checked:’.
- **QA module**: Collect the text followed by ’answer is’ and remove any wrapping quotation marks.
- **AG module**: Turn the truncated response to lower case. Collect the text followed by ’agreement:’ and check if the collected text contains ’true’ or not.

Besides the post-processing, we also formatted the given evidence in our style when preparing the input for the QA module. As demonstrated in examples in Appendix C, we added the tag of ’[(primary/secondary) trial, (section ID) cohort (1/2)]’ on each line of the evidence. We added this tag to prevent LLM from forgetting which trial/cohort this piece of evidence is about.

B Hyperparameters for fine-tuning experiments

For the hyperparameters for fine-tuning experiments, please see Table 2.

Hyperparameter	**Values**
Epochs	3, **5**, 10
Batch size	**8**, 16
Learning rate (LR)	1.0E-5, **2.0E-5**
LR scheduler	**Linear**, Cosine
Weight decay	**0**, 2.0E-7, 1.0E-6
LoRA r	**8**, 16, 32, 64, 128, 256
LoRA α	**16**, 24, 32

Table 2: Hyperparameter for the fine-tuning experiments. Hyperparameters in bold are what we used for the fine-tuned modules reported in Table 1.

C Examples used in prompts

In this section, we list all the examples we used in our prompts for those who wish to reproduce our results. QG examples used for the label generation is in C.1. Rest of QG examples are listed in C.2 $\sim$ C.9, depending on the type and section ID. QA examples are listed in C.10 $\sim$ C.13. AG examples are listed in C.14 $\sim$ C.17.

C.1 QG examples: label generation

The following is the full list of 8 examples used for Prompt 1:

```
You said: The majority of patients in the
    primary trial did not experience Left
    ventricular dysfunction.
```

```
To verify it,
I checked: What is the percentage of patients
     experienced Left ventricular dysfunction?
I checked: What is the percentage of patients
     had any adverse event?

You said: Patients with a QT interval longer
     than half a second are excluded from the
     primary trial.
To verify it,
I checked: How many patients have a QT interval
     longer than half a second?
I checked: What is the maximum QT interval?

You said: Cohort 1 of the primary trial had a
     longer PFS than cohort 2. However the
     patient with the longest PFS was in cohort
     2.
To verify it,
I checked: What is the median PFS of Cohort 1?
I checked: What is the median PFS of Cohort 2?
I checked: What is the longest PFS of Cohort 1?
I checked: What is the longest PFS of Cohort 2?

You said: Japanese participants with an ECOG of
     2 are eligible for the secondary trial and
     the primary trial.
To verify it,
I checked: Are Japanese participants eligible
     for the secondary trial?
I checked: Are Japanese participants eligible
     for the primary trial?
I checked: Are participants with an ECOG of 2
     eligible for the secondary trial?
I checked: Are participants with an ECOG of 2
     eligible for the primary trial?

You said: Less than 1% of patients in the
     primary trial became depressed.
To verify it,
I checked: What is the percentage of patients
     became depressed?
I checked: What is the percentage of patients
     had any adverse event?

You said: There is one case of Cardiopulmonary
     failure in cohort 2 of the secondary trial,
     but none in cohort 1 of the primary trial.
To verify it,
I checked: How many Cardiopulmonary failure are
     there in cohort 2?
I checked: How many Cardiopulmonary failure are
     there in cohort 1?

You said: the primary trial participants apply
     topical imiquimod to cutaneous lesions once
     daily on days for a total of 13 days every
     28 day cycle.
To verify it,
I checked: How many times participants apply
     topical imiquimod to cutaneous lesions daily
     ?
I checked: How many days participants apply
     topical imiquimod to cutaneous lesions
     during every 28 day cycle?

You said: the primary trial administers the
     placebo and Urea/Lactic Acid Cream in the
     same frequency and on the same areas of the
     skin.
To verify it,
I checked: What is the frequency the placebo
     cream is applied?
I checked: What is the frequency the Urea/Lactic
      Acid cream is applied?
I checked: Which areas of the skin the placebo
     cream is applied?
I checked: Which areas of the skin the Urea/
     Lactic Acid cream is applied?
```

C.2 QG examples: results, single

The following is the full list of 9 examples used for custom fine-tuned QG module for `section ID = Results`, `type = Single`:

```
You said: All participants in the primary trial
     had a Central Nervous System (CNS)
     Progression Free Survival (PFS) >= 16 weeks.
To verify it,
I checked: What is the percentage of
     participants with a Central Nervous System (
     CNS) Progression Free Survival (PFS) >= 16
     weeks in the primary trial?

You said: Cohort 1 of the primary trial had a
     longer median, maximum and minimum pfs than
     cohort 2.
To verify it,
I checked: What is the median pfs of cohort 1 in
      the primary trial?
I checked: What is the maximum pfs of cohort 1
     in the primary trial?
I checked: What is the minimum pfs of cohort 1
     in the primary trial?
I checked: What is the median pfs of cohort 2 in
      the primary trial?
I checked: What is the maximum pfs of cohort 2
     in the primary trial?
I checked: What is the minimum pfs of cohort 2
     in the primary trial?

You said: The majority of the primary trial
     subjects either had Progressive disease or
     undetermined CNS objective response rate.
To verify it,
I checked: What is the percentage of subjects
     that had either had progressive disease or
     undetermined CNS objective response rate in
     the primary trial?

You said: The Maximum Tolerated Dose (MTD) of of
      1-Pemetrexed-d-tryptophan (indoximod)
     observed in the primary trial was 800 mg.
To verify it,
I checked: What is the maximum dose of 1-
     Pemetrexed-d-tryptophan (indoximod) observed
      in the primary trial?

You said: In the primary trial 11.1% of patients
      had serious adverse events, no patients had
      serious Adverse Drug Reactions, and over
     half of patients had Unexpected adverse
     events.
To verify it,
I checked: What is the percentage of patients
     that had serious adverse events in the
     primary trial?
```

```
I checked: What is the percentage of patients
     that had serious Adverse Drug Reactions in
     the primary trial?
I checked: What is the percentage of patients
     that had unexpected adverse events?

You said: There was just over 36 hours
     difference in Median Duration of Grade 4
     Neutropenia During Cycle 1 of Chemotherapy
     for the two arms of the primary trial.
To verify it,
I checked: What is the Median Duration of Grade
     4 Neutropenia During Cycle 1 of Chemotherapy
      of cohort 1 in the primary trial?
I checked: What is the Median Duration of Grade
     4 Neutropenia During Cycle 1 of Chemotherapy
      of cohort 2 in the primary trial?

You said: In total more participants in the
     primary trial had no tumor Response, than
     partial response, and only 3 patients had a
     complete response.
To verify it,
I checked: How many patients had no tumor
     Response in the primary trial?
I checked: How many patients had partial
     response in the primary trial?
I checked: How many patients had a complete
     response in the primary trial?

You said: All patients in the Letrozole group of
      the primary trial had a decreased Bone
     Mineral Density of the Lumbar Spine after 3
     years.
To verify it,
I checked: Which cohort is the Letrozole group?
I checked: How many patients in the cohort 1 had
      a decreased Bone Mineral Density of the
     Lumbar Spine after 3 years?
I checked: How many patients in the cohort 2 had
      a decreased Bone Mineral Density of the
     Lumbar Spine after 3 years?

You said: 30% of the primary trial participants
     had an increased level of CECs after 3 weeks
      of Cediranib Maleate treatment.
To verify it,
I checked: What is the percentage of
     participants in the primary trial with an
     increased level of CECs after 3 weeks of
     Cediranib Maleate treatment?
```

C.3 QG examples: results, comparison

The following is the full list of 8 examples used for custom fine-tuned QG module for `section ID = Results`, `type = Comparison`:

```
You said: Patients in the primary trial treated
     with Fulvestrant 250 mg had a 33% higher
     Objective response rate than those treated
     with Fulvestrant 250 mg + Loading Dose.
To verify it,
I checked: Which cohort in the primary trial
     treated patients with Fulvestrant 250 mg?
I checked: Which cohort in the primary trial
     treated patients with Fulvestrant 250 mg +
     Loading Dose?
I checked: What is the objective response rate
     in cohort 1 in the primary trial?
I checked: What is the objective response rate
     in cohort 2 in the primary trial?

You said: the secondary trial and the primary
     trial use similar outcome measures,
     evaluating how long patients survive after
     treatment.
To verify it,
I checked: Does the secondary trial outcome
     measure how long patients survive after
     treatment?
I checked: Does the primary trial outcome
     measure how long patients survive after
     treatment?

You said: participants from both cohorts of the
     primary trial had a drastically lower CNS
     Objective Response Rate than those in the
     secondary trial.
To verify it,
I checked: What is the CNS Objective Response
     Rate of cohort 1 in the primary trial?
I checked: What is the CNS Objective Response
     Rate of cohort 2 in the primary trial?
I checked: What is the CNS Objective Response
     Rate of cohort 1 in the secondary trial?
I checked: What is the CNS Objective Response
     Rate of cohort 2 in the secondary trial?

You said: cohort 1 of the secondary trial had a
     much longer median PFS than cohort 1 of the
     primary trial.
To verify it,
I checked: What is the median PFS of cohort 1 of
      the primary trial?
I checked: What is the median PFS of cohort 1 of
      the secondary trial?

You said: the secondary trial and the primary
     trial both use Pathologic complete response
     (pCR) as their outcome measure, and use a 6
     month time frame.
To verify it,
I checked: Is Pathologic complete response (pCR)
      used as the outcome measure in the
     secondary trial?
I checked: Is Pathologic complete response (pCR)
      used as the outcome measure in the primary
     trial?
I checked: Does the secondary trial use a 6
     month time frame?
I checked: Does the primary trial use a 6 month
     time frame?

You said: Both cohorts in the primary trial
     outperformed cohort 1 of the secondary trial
      in median PFS.
To verify it,
I checked: What is the median PFS of cohort 1 of
      the primary trial?
I checked: What is the median PFS of cohort 2 of
      the primary trial?
I checked: What is the median PFS of cohort 1 of
      the secondary trial?

You said: In the primary trial, Dasatinib, 70 mg,
      Twice Daily results in a better median PFS
     than Dasatinib, 100 mg, Daily. The opposite
```

```
    was true in the secondary trial.
To verify it,
I checked: Which cohort in the primary trial
    uses Dasatinib, 70 mg, Twice Daily?
I checked: Which cohort in the primary trial
    uses Dasatinib, 100 mg, Daily?
I checked: What is the median PFS of cohort 1 of
     the primary trial?
I checked: What is the median PFS of cohort 2 of
     the primary trial?
I checked: Which cohort in the secondary trial
    uses Dasatinib, 70 mg, Twice Daily?
I checked: Which cohort in the secondary trial
    uses Dasatinib, 100 mg, Daily?
I checked: What is the median PFS of cohort 1 of
     the secondary trial?
I checked: What is the median PFS of cohort 2 of
     the secondary trial?

You said: More than half the participants of the
     primary trial are considered to be censored
    . the secondary trial used the same outcome
    measurement, but had no censored patients.
To verify it,
I checked: How many participants are there in
    the primary trial?
I checked: How many participants are censored in
     the primary trial?
I checked: Is the secondary trial used the same
    outcome measurement as the primary trial?
I checked: How many participants are censored in
     the secondary trial?
```

C.4 QG examples: intervention, single

The following is the full list of 9 examples used for custom fine-tuned QG module for `section ID = Intervention`, `type = Single`:

```
You said: Intervention of Cohort B is described
    as Afatinib 50 mg, taken orally, for every
    day of the study.
To verify it,
I checked: Is the intervention of Cohort B
    described as Afatinib 50 mg, taken orally,
    for every day of the study?

You said: Patients taking Intervention 1 of the
    primary trial receive 4 times as much
    simvastatin as anastrozole.
To verify it,
I checked: How much simvastatin do patients in
    cohort 1 receive?
I checked: How much anastrozole do patients in
    cohort 1 receive?

You said: All the primary trial participants
    have the same number of calories in their
    diets throughout the duration of the study.
To verify it,
I checked: Does the participants in the primary
    trial have the same number of calories in
    their diets throughout the duration of the
    study?
I checked: What is the percentage of
    participants in the primary trial which have
     the same number of calories in their diets
    throughout the duration of the study?

You said: The difference between the two cohorts
     of the primary trial is that cohort 1
    participated in a Hatha yoga, whereas cohort
     2 abstained from yoga.
To verify it,
I checked: Did cohort 1 of the primary trial
    participated in a Hatha yoga?
I checked: Was cohort 2 of the primary trial
    abstrained from yoga?

You said: Patients in cohort 1 of the primary
    trial may receive gradually increasing doses
     of Afatinib monotherapy, up to 160% of the
    starting dose.
To verify it,
I checked: Do patients in cohort 1 of the
    primary trial receive gradually increasing
    doses of Afatinib monotherapy?
I checked: What is the starting dose of Afatinib
     monotherapy for the cohort of the primary
    trial?
I checked: What is the maximum dose of Afatinib
    monotherapy for the cohort of the primary
    trial?

You said: Participants in group 2 of the primary
     trial receive taping to anastomosis regions
    ., but no Complex Decongestive Physiotherapy
    .
To verify it,
I checked: Do participants in cohort 2 of the
    primary trial receive taping to anastomosis
    regions?
I checked: Do participants in cohort 2 of the
    primary trial receive Complex Decongestive
    Physiotherapy?

You said: Only one cohort of the primary trial
    needs to receive manual lymph drainage prior
     to each cycle of anthracyclines.
To verify it,
I checked: Does cohort 1 of the primary trial
    need to receive manual lymph drainage prior
    to each cycle of anthracyclines?
I checked: Does cohort 2 of the primary trial
    need to receive manual lymph drainage prior
    to each cycle of anthracyclines?

You said: the primary trial does not specificy
    the route of administration of its
    intervention.
To verify it,
I checked: Does the primary trial specify the
    route of administration of its intervention?

You said: Only cohort 2 of the primary trial
    receive letrozole, but both cohorts undergo
    Endocrine Therapy.
To verify it,
I checked: Does cohort 1 of the primary trial
    receive letrozole?
I checked: Does cohort 2 of the primary trial
    receive letrozole?
I checked: Does cohort 1 of the primary trial
    undergo Endocrine Therapy?
I checked: Does cohort 2 of the primary trial
    undergo Endocrine Therapy?
```

C.5 QG examples: intervention, comparison

The following is the full list of 7 examples used for custom fine-tuned QG module for `section ID = Intervention`, `type = Comparison`:

```
You said: Between all cohorts in the primary
    trial and the secondary trial Omega-3-fatty
    Acids are only used in one cohort.
To verify it,
I checked: Is Omega-3-fatty acids used in cohort
     1 in the primary trial?
I checked: Is Omega-3-fatty acids used in cohort
     2 in the primary trial?
I checked: Is Omega-3-fatty acids used in cohort
     1 in the secondary trial?
I checked: Is Omega-3-fatty acids used in cohort
     2 in the secondary trial?

You said: The intervention in the primary trial
    requires patients to receive multiple
    applications of treatment over a period of
    several weeks, whereas the interventions for
     the secondary trial are only apllied once.
To verify it,
I checked: Does the intervention in cohort 1 in
    the primary trial require patients to
    receive multiple applications of treatment
    over a period of several weeks?
I checked: Does the intervention in cohort 2 in
    the primary trial require patients to
    receive multiple applications of treatment
    over a period of several weeks?
I checked: Is the intervention in cohort 1 in
    the secondary trial only applied once?
I checked: Is the intervention in cohort 2 in
    the secondary trial only applied once?

You said: Cohort 2 of the secondary trial and
    the primary trial are control groups.
To verify it,
I checked: Is cohort 2 a control group in the
    secondary trial?
I checked: Is cohort 2 a control group in the
    primary trial?

You said: Patients in the primary trial receive
    an Infusion of 3 ml Perflutren Lipid
    Microspheres at a rate of approximately 3ml/
    min, whereas in the secondary trial subjects
     are implanted with a permanent Meso
    BioMatrix Acellular Peritoneum Matrix.
To verify it,
I checked: Do patients in cohort 1 in the
    primary trial receive an Infusion of 3 ml
    Perflutren Lipid Microspheres at a rate of
    approximately 3ml/min?
I checked: Do patients in cohort 2 in the
    primary trial receive an Infusion of 3 ml
    Perflutren Lipid Microspheres at a rate of
    approximately 3ml/min?
I checked: Are trial subjects in cohort 1 in the
     secondary trial implanted with a permanent
    Meso BioMatrix Acellular Peritoneum Matrix?
I checked: Are trial subjects in cohort 2 in the
     secondary trial implanted with a permanent
    Meso BioMatrix Acellular Peritoneum Matrix?

You said: All Participants in the primary trial
    and the secondary trial are receiving the
    same daily dose of Degarelix injected into
    the muscle, for the same duration of time.
To verify it,
I checked: What is the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 1 in the primary
    trial?
I checked: What is the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 2 in the primary
    trial?
I checked: What is the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 1 in the secondary
    trial?
I checked: What is the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 2 in the secondary
    trial?
I checked: What is the duration of time that
    Degarelix injected into the muscle for the
    participants in cohort 1 in the primary
    trial?
I checked: What is the duration of time that
    Degarelix injected into the muscle for the
    participants in cohort 2 in the primary
    trial?
I checked: What is the duration of time that
    Degarelix injected into the muscle for the
    participants in cohort 1 in the secondary
    trial?
I checked: What is the duration of time that
    Degarelix injected into the muscle for the
    participants in cohort 2 in the secondary
    trial?

You said: Each patient in the primary trial
    receives 3 different drugs, whereas in the
    secondary trial patients receive 1 of 2
    possible drugs.
To verify it,
I checked: How many different drugs each patient
     in cohort 1 in the primary trial receive?
I checked: How many different drugs each patient
     in cohort 2 in the primary trial receive?
I checked: How many different drugs each patient
     in cohort 1 in the secondary trial receive?
I checked: How many different drugs each patient
     in cohort 2 in the secondary trial receive?

You said: Breast breast irradiation is used in
    some form for both cohorts of the secondary
    trial, but not at all in the primary trial.
To verify it,
I checked: Is breast irradiation used in some
    form for cohort 1 of the secondary trial?
I checked: Is breast irradiation used in some
    form for cohort 2 of the secondary trial?
I checked: Is breast irradiation used in some
    form for cohort 1 of the primary trial?
I checked: Is breast irradiation used in some
    form for cohort 2 of the secondary trial?
```

C.6 QG examples: adverse events, single

The following is the full list of 9 examples used for custom fine-tuned QG module for `section ID = Adverse Events`, `type = Single`:

```
You said: There were 2 cases of severe back pain
     observed in the primary trial.
To verify it,
I checked: How many cases of severe back pain
    were observed in the primary trial?

You said: The same number of cases of
    Neutropenia, Febrile neutropenia and
    Pancytopenia are observed in patients from
    cohort 1 of the primary trial.
To verify it,
I checked: How many patients from cohort 1 of
    the primary trial had Neutropenia?
I checked: How many patients from cohort 1 of
    the primary trial had Febrile neutropenia?
I checked: How many patients from cohort 1 of
    the primary trial had Pancytopenia?

You said: There were 2 more cases of
    Gastrointestinal Haemorrhage in cohort 2 of
    the primary trial, than in cohort 1.
To verify it,
I checked: How many cases of Gastrointestinal
    Haemorrhage are there in cohort 1?
I checked: How many cases of Gastrointestinal
    Haemorrhage are there in cohort 2?

You said: 80% of patients in the primary trial
    did not suffer any adverse events.
To verify it,
I checked: What is the percentage of patients in
     the primary trial did not suffer any
    adverse event?

You said: 3/112 patients (2.68%) in the primary
    trial had Diabetes insipidus.
To verify it,
I checked: What is the percentage of patients
    had diabetes insipidus?

You said: Cohort 1 of the primary trial had a
    much higher number of deaths than cohort 2.
To verify it,
I checked: How many deaths were there in cohort
    1 of the primary trial?
I checked: How many deaths were there in cohort
    2 of the primary trial?

You said: Cohort 1 of the primary trial recorded
     no deaths and no cases of Enterocolitis
    infectious.
To verify it,
I checked: How many deaths were there in cohort
    1 of the primary trial?
I checked: How many cases of Enterocolitis
    infectious were there in cohort 1 of the
    primary trial?

You said: At least one patient in the primary
    trial suffered from a life threatening bone
    fracture.
To verify it,
I checked: How many patients in the primary
    trial suffered from a life threatening bone
    fracture?

You said: In the primary trial, all cases of
    Enteritis, Vertigo and Anaemia occurred in
    cohort 2.
To verify it,
I checked: How many cases of Enteritis occurred
    in cohort 1 in the primary trial?
I checked: How many cases of Vertigo occurred in
     cohort 1 in the primary trial?
I checked: How many cases of Anaemia occurred in
     cohort 1 in the primary trial?
I checked: How many cases of Enteritis occurred
    in cohort 2 in the primary trial?
I checked: How many cases of Vertigo occurred in
     cohort 2 in the primary trial?
I checked: How many cases of Anaemia occurred in
     cohort 2 in the primary trial?
```

C.7 QG examples: adverse events, comparison

The following is the full list of 9 examples used for custom fine-tuned QG module for `section ID = Adverse Events`, `type = Comparison`:

```
You said: the secondary trial recorded more
    total occurences of gastrointestinal adverse
     events than the primary trial.
To verify it,
I checked: How many total occurences of
    gastrointestinal adverse events were
    recorded in the primary trial?
I checked: How many total occurences of
    gastrointestinal adverse events were
    recorded in the secondary trial?

You said: the primary trial had a lower total
    percentage of patients experiencing adverse
    events compared to the secondary trial.
To verify it,
I checked: What is the percentage of patients
    experienced adverse events in the primary
    trial?
I checked: What is the percentage of patients
    experienced adverse events in the secondary
    trial?

You said: Some of the patients in the primary
    trial were recorded as having heart related
    adverse events, whereas many patients in the
     secondary trial experienced several
    different breathing related issues.
To verify it,
I checked: How many patients recorded as heart
    related adverse events in the primary trial?
I checked: How many patients experienced
    breathing related issues recorded in the
    secondary trial?

You said: Between both of the patient cohort of
    the primary trial and the secondary trial
    there was only a single patient who suffered
     heart failure.
To verify it,
I checked: How many patients suffered with heart
     failure in the primary trial?
I checked: How many patients suffered with heart
     failure in the secondary trial?

You said: the primary trial and the secondary
    trial recorded the same proportion of
    patients experiencing nausea.
```

```
To verify it,
I checked: What is the proportion of patients
     experienced nausea in the primary trial?
I checked: What is the proportion of patients
     experienced nausea in the secondary trial?

You said: the primary trial had three times the
     occurence rate of fistula enterovesical as
     the secondary trial.
To verify it,
I checked: What is the occurence rate of fistula
      enterovesical in the primary trial?
I checked: What is the occurence rate of fistula
      enterovesical in the secondary trial?

You said: There were no completed suicides in
     either the primary trial or the secondary
     trial, however there was one attempt in
     cohort 1 of the secondary trial.
To verify it,
I checked: How many completed suicides were
     there in the primary trial?
I checked: How many completed suicides were
     there in the secondary trial?
I checked: How many attempted suicides were
     there in the secondary trial?

You said: Heart-related adverse events were
     recorded in the secondary trial, but not the
      primary trial.
To verify it,
I checked: How many heart-related adverse events
      were recorded in the secondary trial?
I checked: How many heart-related adverse events
      were recorded in the primary trial?

You said: The adverse events in the primary
     trial where all equally prevalent, whereas
     in the secondary trial, alcohol poisoning
     was reported as the most common event.
To verify it,
I checked: Are all the adverse events equally
     prevalent in the primary trial?
I checked: Is alcohol poisoning the most common
     adverse event in the secondary trial?
```

C.8 QG examples: eligibility, single

The following is the full list of 9 examples used for custom fine-tuned QG module for `section ID = Eligibility`, `type = Single`:

```
You said: Paula recently had fell down a flight
     of stairs and fractured her hip, she is
     excluded from the primary trial.
To verify it,
I checked: Are female patients excluded from the
      primary trial?
I checked: Are female patients with fractured
     hip excluded from the primary trial?

You said: Spanish women with a heart rate of at
     least 60 beats per minute are eligible for
     the primary trial.
To verify it,
I checked: Are Spanish women eligible for the
     primary trial?
I checked: Are patients with a heart rate of at
     least 60 beats per minute eligible for the
     primary trial?

You said: Patients must have a life expectancy
     over a year to participate in the primary
     trial.
To verify it,
I checked: What is the minimum life expectancy
     for the primary trial?

You said: Patients must have a white blood cell
     count above 1,500/mm3 to participate in the
     primary trial.
To verify it,
I checked: What is the minimum blood cell count
     for the primary trial?

You said: A patient who had a Joint injection in
      the last month would not be eligible for
     the primary trial.
To verify it,
I checked: Are patients who had a Joint
     injection in the last month eligible for the
      primary trial?

You said: Abnormal LVEF, Pregnancy or lactating
     automatically eliminates patients from
     participating in the primary trial.
To verify it,
I checked: Are patients with Abnormal LVEF
     excluded from the primary trial?
I checked: Are patients with Pregnancy excluded
     from the primary trial?
I checked: Are patients with lactating excluded
     from the primary trial?

You said: Patients prescribed with bisoprolol or
      labetalol to treat CHF are eligible for the
      primary trial.
To verify it,
I checked: Are patients prescribed with
     bisoprolol to treat CHF eligible for the
     primary trial?
I checked: Are patients prescribed with
     labetalol to treat CHF eligible for the
     primary trial?

You said: Participants for the primary trial
     must be in pairs, a breast cancer survivor
     and a caregiver, both must either be fluent
     in english or spanish.
To verify it,
I checked: Are participants for the primary
     trial must be in pairs, a breast cancer
     survivor and a caregiver?
I checked: What language must breast cancer
     survivor be fluent in?
I checked: What language must caregivers be
     fluent in?

You said: A patient who underwent T-cell
     transfer therapy in the past 6 months prior,
      and has fully recovered from the associated
      toxicities, would be excluded from the
     primary trial.
To verify it,
I checked: Are patients who underwent T-cell
     transfer therapy in the past 6 months prior
     excluded from the primary trial?
I checked: Are patients who recovered from the
     associated toxicities excluded from the
     primary trial?
```

C.9 QG examples: eligibility, comparison

The following is the full list of 8 examples used for custom fine-tuned QG module for `section ID = Eligibility`, `type = Comparison`:

```
You said: Patients suffering from vomiting are
     still eligible for both the secondary trial
     and the primary trial.
To verify it,
I checked: Are patients suffering from vomiting
     eligible for the secondary trial?
I checked: Are patients suffering from vomiting
     eligible for the primary trial?

You said: Patients with cytologically confirmed
     breast cancer, who’s Locally recurrent
     disease is amenable to radiation with
     curative intent are not eligible for the
     secondary trial, but are eligible for the
     primary trial.
To verify it,
I checked: Are patients with cytologically
     confirmed breast cancer eligible for the
     secondary trial?
I checked: Are patients with cytologically
     confirmed breast cancer eligible for the
     primary trial?
I checked: Are patients whose locally recurrent
     disease is amenable to radiation with
     curative intent eligible for the secondary
     trial?
I checked: Are patients whose locally recurrent
     disease is amenable to radiation with
     curative intent eligible for the primary
     trial?

You said: the secondary trial and the primary
     trial both accept patients with HER-2/neu-
     overexpressing adenocarcinoma of the breast.
To verify it,
I checked: Are patients with HER-2/neu-
     overexpressing adenocarcinoma of the breast
     eligible for the secondary trial?
I checked: Are patients with HER-2/neu-
     overexpressing adenocarcinoma of the breast
     eligible for the primary trial?

You said: Agatha had her 53rd birthday last week,
      she has a histologically confirmed
     adenocarcinoma of the breast, with no
     evidence of metastatic disease. Agatha is of
      white british origin. she is eligible for
     the primary trial but not the secondary
     trial, due to her age.
To verify it,
I checked: Are women of 53 years old eligible
     for the primary trial?
I checked: Are women of 53 years old eligible
     for the secondary trial?
I checked: Are patients with a histologically
     confirmed adenocarcinoma of the breast with
     no evidence of metastatic disease eligible
     for the primary trial?
I checked: Are patients of white british origin
     eligible for the primary trial?
You said: Completely disabled patients below the
      age of 20, totally confined to bed or chair
      and unable to carry on any selfcare are
     eligible for the primary trial but excluded
     from the secondary trial.
To verify it,
I checked: Are patients completely disabled
     eligible for the primary trial?
I checked: Are patients completely disabled
     eligible for the secondary trial?
I checked: Are patients below the age of 20
     eligible for the primary trial?
I checked: Are patients below the age of 20
     eligible for the secondary trial?
I checked: Are patients totally confined to bed
     or chair and unable to carry on any selfcare
      eligible for the primary trial?
I checked: Are patients totally confined to bed
     or chair and unable to carry on any selfcare
      eligible for the secondary trial?

You said: Only White and Asian patients are
     eligible for both the primary trial and the
     secondary trial.
To verify it,
I checked: Are patients other than White or
     Asian excluded from the primary trial?
I checked: Are patients other than White or
     Asian excluded from the secondary trial?

You said: Women of any age can participate in
     the primary trial or the secondary trial.
To verify it,
I checked: Are women of any age eligible for the
      primary trial?
I checked: Are women of any age eligible for the
      secondary trial?

You said: All cancer stages are accepted for the
      primary trial and the secondary trial.
To verify it,
I checked: Are patients with any cancer stage
     eligible for the primary trial?
I checked: Are patients with any cancer stage
     eligible for the secondary trial?
```

C.10 QA examples: results

The following is the full list of 3 examples used for custom fine-tuned QA module for `section ID = Results`:

```
#########################################
*** Question: What is the maximum pfs of cohort
     2 in the primary trial?
*** Evidence:
[primary trial, Outcome Measurement] Progression
      Free Survival
[primary trial, Outcome Measurement] Time frame:
      From the date of randomization to date of
     first documented disease progression,
     assessed up to 39 months.
[primary trial, Results cohort 1] Arm/Group
     Title: A (Sorafenib + Gemcitabine or
     Capecitabine)
[primary trial, Results cohort 1] Arm/Group
     Description: Sorafenib will be administered
     (400 mg; 2 tablets x 200 mg) orally twice
     daily
```

```
[primary trial, Results cohort 1] Gemcitabine: Gemcitabine will be administered 1000 mg/m2 pm Days 1 and 8 of a 21 day cycle
[primary trial, Results cohort 1] Sorafenib: Sorafenib will be administered (400 mg; 2 tablets x 200 mg) orally twice daily
[primary trial, Results cohort 1] Capecitabine: Capecitabine will be administered orally at a dose of 1,000 mg/m2 twice daily
[primary trial, Results cohort 1] Overall Number of Participants Analyzed: 81
[primary trial, Results cohort 1] Median (95% Confidence Interval)
[primary trial, Results cohort 1] Unit of Measure: Days 103 (83 to 128)
[primary trial, Results cohort 2] Arm/Group Title: B (Placebo + Gemcitabine or Capecitabine)
[primary trial, Results cohort 2] Arm/Group Description: Placebo will be administered ( 2 tablets ) orally twice daily
[primary trial, Results cohort 2] Gemcitabine: Gemcitabine will be administered 1000 mg/m2 pm Days 1 and 8 of a 21 day cycle
[primary trial, Results cohort 2] Placebo: Placebo will be administered (400 mg; 2 tablets x 200 mg) orally twice daily
[primary trial, Results cohort 2] Capecitabine: Capecitabine will be administered orally at a dose of 1,000 mg/m2 twice daily
[primary trial, Results cohort 2] Overall Number of Participants Analyzed: 79
[primary trial, Results cohort 2] Median (95% Confidence Interval)
[primary trial, Results cohort 2] Unit of Measure: Days 81 (48 to 95)
*** Answer:
Let’s think step by step. From evidence "[primary trial, Outcome Measurement] Progression Free Survival"
and "[primary trial, Results cohort 2] Median (95% Confidence Interval)"
and "[primary trial, Results cohort 2] Unit of Measure: Days 81 (48 to 95)",
maximum pfs(Progression Free Survival) of cohort 2 in the primary trial is 95 Days.
Therefore, answer is "95 Days"

#########################################
*** Question: What is the median PFS of cohort 1 of the secondary trial?
*** Evidence:
[primary trial, Outcome Measurement] Number of Participants With Progression Free Survival (PFS) in HER2
[primary trial, Outcome Measurement] Time frame: From the date of randomization until the date of the first documented progression or date of death from any cause
[primary trial, Results cohort 1] Arm/Group Title: Placebo + Letrozole 2.5 mg
[primary trial, Results cohort 1] Arm/Group Description: Participants received 6 tablets of placebo and 1 tablet of letrozole 2.5 milligrams (mg) orally daily.
[primary trial, Results cohort 1] Overall Number of Participants Analyzed: 108
[primary trial, Results cohort 1] Measure Type: Count of Participants
[primary trial, Results cohort 1] Unit of Measure: Participants 89 82.4%
[primary trial, Results cohort 2] Arm/Group Title: Lapatinib 1500 mg + Letrozole 2.5 mg
[primary trial, Results cohort 2] Arm/Group Description: Participants received 6 tablets of Lapatinib orally daily
[primary trial, Results cohort 2] Overall Number of Participants Analyzed: 111
[primary trial, Results cohort 2] Measure Type: Count of Participants
[primary trial, Results cohort 2] Unit of Measure: Participants 88 79.3%
[secondary trial, Outcome Measurement] Time to Disease Progression (Initial Treatment)
[secondary trial, Outcome Measurement] Time frame: Randomization date to the earliest date of first documented disease progression date or the date of death if the participant died due to study disease (up to 82 months)
[secondary trial, Results cohort 1] Arm/Group Title: Gemcitabine Plus Docetaxel
[secondary trial, Results cohort 1] Arm/Group Description: gemcitabine 1000 milligrams plus docetaxel 75 mg/m2, intravenous.
[secondary trial, Results cohort 1] Overall Number of Participants Analyzed: 239
[secondary trial, Results cohort 1] Median (95% Confidence Interval)
[secondary trial, Results cohort 1] Unit of Measure: months 9.28 (7.73 to 10.79)
[secondary trial, Results cohort 2] Arm/Group Title: Docetaxel Plus Capecitabine
[secondary trial, Results cohort 2] Arm/Group Description: docetaxel 75 mg/m2, intravenous , day 1 every 21 days plus capecitabine 1000 mg/m2
[secondary trial, Results cohort 2] Overall Number of Participants Analyzed: 236
[secondary trial, Results cohort 2] Median (95% Confidence Interval)
[secondary trial, Results cohort 2] Unit of Measure: months 8.88 (7.37 to 11.05)
*** Answer:
Let’s think step by step. Among all evidence related to the cohort 1 of the secondary trial,
there is no mention of median PFS.
Therefore, answer is "Not Available"

#########################################
*** Question: Which cohort in the primary trial uses Dasatinib, 70 mg, Twice Daily?
*** Evidence:
[primary trial, Outcome Measurement] Progression-free Survival
[primary trial, Outcome Measurement] RECIST progression defined as 20% increase in the sum of longest diameters of target measurable lesions
[primary trial, Outcome Measurement] Time frame: Up to 2 years
[primary trial, Results cohort 1] Arm/Group Title: Dasatinib, 100 mg, Daily
[primary trial, Results cohort 1] Arm/Group Description: Dasatinib, 100 mg PO daily until progression of disease
[primary trial, Results cohort 1] Overall Number of Participants Analyzed: 41
[primary trial, Results cohort 1] Median (95%
    Confidence Interval)
[primary trial, Results cohort 1] Unit of
    Measure: weeks 10.3 (8.4 to 16.7)
[primary trial, Results cohort 2] Arm/Group
    Title: Dasatinib, 70 mg, Twice Daily
[primary trial, Results cohort 2] Arm/Group
    Description: Dasatinib, 70 mg PO twice daily
     until progression of disease
[primary trial, Results cohort 2] Overall Number
     of Participants Analyzed: 38
[primary trial, Results cohort 2] Median (95%
    Confidence Interval)
[primary trial, Results cohort 2] Unit of
    Measure: weeks 15.3 (8.7 to 20.1)
[secondary trial, Outcome Measurement] Number of
     Patients With Objective Treatment Response
    (Complete or Partial) in the CNS
[secondary trial, Outcome Measurement] Imaging
    was performed at 8-week intervals to assess
    response to treatment.
[secondary trial, Outcome Measurement] Time
    frame: Baseline scan prior to study entry
    was performed within 14 days of cycle 1 day
    1
[secondary trial, Results cohort 1] Arm/Group
    Title: Irinotecan and Temozolomide
[secondary trial, Results cohort 1] Arm/Group
    Description: irinotecan hydrochloride
    administered intravenously (IV)
[secondary trial, Results cohort 1] temozolomide
     orally for seven days at a starting dose of
     100 mg/m2 on days 1-7 and days 15-21 of a
    28 day cycle
[secondary trial, Results cohort 1] Overall
    Number of Participants Analyzed: 30
[secondary trial, Results cohort 1] Measure Type:
     Count of Participants
[secondary trial, Results cohort 1] Unit of
    Measure: Participants 2 6.7%
*** Answer:
Let's think step by step. From evidence "[
    primary trial, Results cohort 2] Arm/Group
    Title: Dasatinib, 70 mg, Twice Daily",
cohort 2 in the primary trial uses Dasatinib, 70
     mg, twice Daily.
Therefore, answer is "cohort 2"
```

C.11 QA examples: intervention

The following is the full list of 3 examples used for custom fine-tuned QA module for `section ID = Intervention`:

```
#########################################
*** Question: Does the participants in the
    primary trial have the same number of
    calories in their diets throughout the
    duration of the study?
*** Evidence:
[primary trial, INTERVENTION cohort 1]
    Behavioral Dietary Intervention
[primary trial, INTERVENTION cohort 1] Beginning
     2-4 weeks after completion of lumpectomy,
    patients receive food diaries to complete
    for 7-10 days. Dietary counselors then give
    patients guidelines for dietary
    modifications to reduce caloric intake by
    25% of their normal diet. Patients follow
    caloric restricted diet for 10 weeks (2
    weeks prior to radiation therapy, during 6
    weeks of radiation therapy, and at least 2
    weeks after radiation therapy). Patients
    undergo radiation therapy QD 5 days a week
    for 6 weeks.
[primary trial, INTERVENTION cohort 1]
    Behavioral dietary intervention: Receive
    caloric restricted dietary intervention
[primary trial, INTERVENTION cohort 1]
    Therapeutic conventional surgery: Undergo
    definitive lumpectomy
[primary trial, INTERVENTION cohort 1] Radiation
     therapy: Undergo radiation therapy
[primary trial, INTERVENTION cohort 1]
    Counseling intervention: Receive dietary
    counseling
[primary trial, INTERVENTION cohort 1] Quality-
    of-life assessment: Ancillary studies
*** Answer:
Let's think step by step. From evidence "[
    primary trial, INTERVENTION cohort 1]
    Beginning 2-4 weeks after completion of
    lumpectomy, patients receive food diaries to
     complete for 7-10 days. Dietary counselors
    then give patients guidelines for dietary
    modifications to reduce caloric intake by
    25% of their normal diet. Patients follow
    caloric restricted diet for 10 weeks (2
    weeks prior to radiation therapy, during 6
    weeks of radiation therapy, and at least 2
    weeks after radiation therapy). Patients
    undergo radiation therapy QD 5 days a week
    for 6 weeks.",
patients reduce caloric intake by 25% of their
    normal diet beginning 2-4 weeks after
    completion of lupectomy,
which means the number of calories in their
    diets change throughout the duration of the
    study.
Therefore, answer is "No"

#########################################
*** Question: How many different drugs each
    patient in cohort 2 in the primary trial
    receive?
*** Evidence:
[primary trial, INTERVENTION 1] Pertuzumab +
    Trastuzumab + Taxane
[primary trial, INTERVENTION 1] Participants
    received pertuzumab and trastuzumab
    intravenously (IV) plus taxane chemotherapy
    once of every 3 weeks per treatment cycle
    until predefined study end, unacceptable
    toxicity, withdrawal of consent, disease
    progression, or death, whichever occurred
    first. Taxane chemotherapy was docetaxel,
    paclitaxel, or nab-paclitaxel, per the
    investigator's choice.[secondary trial,
    INTERVENTION 1] Afatinib 50 mg
[secondary trial, INTERVENTION 1] Patients
    received continuous daily dosing with
    Afatinib 50 mg orally from Day 1 to Day 21
    of each treatment course. 2 treatment
    courses were to be given in the trial.
[secondary trial, INTERVENTION 2] Lapatinib 1500
     mg
[secondary trial, INTERVENTION 2] Patients
    received continuous daily dosing with
    Lapatinib 1500 mg orally from Day 1 to Day
    21 of each treatment course. 2 treatment
    courses were to be given in the trial.
*** Answer:
Let’s think step by step. Among all evidence,
there is no mention of cohort 2 in the primary
    trial.
Therefore, answer is "Not available"

#########################################
*** Question: What is the daily dose of
    Degarelix injected into the muscle for the
    participants in cohort 2 in the primary
    trial?
*** Evidence:
[primary trial, INTERVENTION cohort 1] Arm A:
    Triptorelin + Letrozol
[primary trial, INTERVENTION cohort 1] Arm A:
    Triptorelin 3.75 mg i.m. on day 1 every 28
    days for 6 cycles + letrozole 2.5 mg/day
    orally for 6 cycles
[primary trial, INTERVENTION cohort 1]
    Triptorelin: Triptorelin 3.75 mg injected
    into the muscle on day 1 every 28 days for 6
     cycles (1 cycle= 28 days)
[primary trial, INTERVENTION cohort 1] Letrozole:
     Letrozole 2.5 mg orally every day for 6
    cycles
[primary trial, INTERVENTION cohort 2] Arm B:
    Degarelix + Letrozol
[primary trial, INTERVENTION cohort 2] Arm B:
    Degarelix 240 mg s.c. on day 1 of cycle 1,
    followed by 80 mg s.c. on day 1 of cycles 2
    to 6 + letrozole 2.5 mg every day orally for
     6 cycles
[primary trial, INTERVENTION cohort 2] Degarelix:
     Degarelix 240 mg injected under the skin
    given as two injections of 120 mg on the
    first day of treatment, followed by
    injection of 80 mg on day 1 of cycles 2 to 6
     (1 cycle=28 days)
[primary trial, INTERVENTION cohort 2] Letrozole:
     Letrozole 2.5 mg orally every day for 6
    cycles[secondary trial, INTERVENTION 1]
    Zoledronic Acid Upfront
[secondary trial, INTERVENTION cohort 1]
    Participants in the upfront arm received
    Zoledronic Acid 4 mg i.v. on Day 1 and every
     6 months until disease progression (
    recurrence)or the end of study. Participants
     also received Letrozole 2.5 daily plus
    calcium (1000-1200 mg) and vitamin D
    (400-800 IU) daily.
[secondary trial, INTERVENTION cohort 1]
    Letrozole : Participants received 2.5 mg
    daily.
[secondary trial, INTERVENTION cohort 1]
    Zoledronic Acid : Participants received
    Zoledronic Acid 4 mg IV 15-minute infusion
    every 6 months.
[secondary trial, INTERVENTION cohort 2]
    Zoledronic Acid Delayed-start
[secondary trial, INTERVENTION cohort 2] In lieu
     of a placebo arm, which was considered
    unethical for this trial, a delayed start
    arm was used. Participants who met certain
    clinical criteria indicating risk of lumbar
    spine or total hip fracture, or experienced
    clinical fracture unrelated to trauma or any
     asymptomatic fracture discovered at the
    Month 36 scheduled visit, were started on
    zoledronic acid 4 mg i.v. and for every 6
    months until disease progression (recurrence
    ) or end of study. Participants also
    received Letrozole 2.5 daily plus calcium
    (1000-1200 mg) and vitamin D (400-800 IU)
    daily.
[secondary trial, INTERVENTION cohort 2]
    Letrozole : Participants received 2.5 mg
    daily.
[secondary trial, INTERVENTION cohort 2]
    Zoledronic Acid : Participants received
    Zoledronic Acid 4 mg IV 15-minute infusion
    every 6 months.
*** Answer:
Let’s think step by step. From evidence "[
    primary trial, INTERVENTION cohort 2]
    Degarelix: Degarelix 240 mg injected under
    the skin given as two injections of 120 mg
    on the first day of treatment, followed by
    injection of 80 mg on day 1 of cycles 2 to 6
     (1 cycle=28 days)",
daily dose of Degarelix is 240 mg on the first
    day of treatment, followed by injection of
    80 mg on day 1 of cycles 2 to 6 (1 cycle=28
    days).
Therefore, answer is "240 mg on the first day of
     treatment, followed by injection of 80 mg
    on day 1 of cycles 2 to 6 (1 cycle=28 days)"
```

C.12 QA examples: adverse events

The following is the full list of 3 examples used for custom fine-tuned QA module for `section ID = Adverse Events`:

```
#########################################
*** Question: What is the percentage of patients
     in the primary trial did not suffer any
    adverse event?
*** Evidence:
[primary trial, Adverse Events cohort 1] Total:
    1/4 (25.00%)
[primary trial, Adverse Events cohort 1] Vertigo
     0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Abdominal adhesions 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Abdominal distension 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Abdominal pain 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Diarrhoea 0/4 (0.00%)
[primary trial, Adverse Events cohort 1] Nausea
    0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Vomiting 0/4 (0.00%)
[primary trial, Adverse Events cohort 1] Disease
     progression 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Influenza 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Nasopharyngitis 0/4 (0.00%)
[primary trial, Adverse Events cohort 1] Lumbar
    vertebral fracture 0/4 (0.00%)
[primary trial, Adverse Events cohort 1]
    Hyponatraemia 0/4 (0.00%)
[primary trial, Adverse Events cohort 1] Ataxia
    0/4 (0.00%)
*** Answer:
```

```
Let's think step by step. From evidence "[primary trial, Adverse Events cohort 1] Total: 1/4 (25.00%)",
the percentage of patients in the primary trial with any adverse event is 25.00%.
Then the percentage of patients who did not suffer any adverse event is 100% - 25.00% = 75.00%.
Therefore, answer is "75.00%"

#########################################
*** Question: How many patients experienced breathing related issues recorded in the secondary trial?
*** Evidence:
primary trial, Adverse Events cohort 1] Total: 6/81 (7.41%)
[primary trial, Adverse Events cohort 1] Colitis [1]1/81 (1.23%)
[primary trial, Adverse Events cohort 1] Multiple Sclerosis Relapse 1/81 (1.23%)
[primary trial, Adverse Events cohort 1] Neurotoxicity [2]2/81 (2.47%)
[primary trial, Adverse Events cohort 1] Community-acquired pneumonia 1/81 (1.23%)
[primary trial, Adverse Events cohort 1] Local Infection Reservoir Area 1/81 (1.23%)
[secondary trial, Adverse Events cohort 1] Total: 20/52 (38.46%)
[secondary trial, Adverse Events cohort 1] Anaemia 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Pancytopenia 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Acute myocardial infarction 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Atrial fibrillation 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Cardiac failure 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Cardiogenic shock 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Palpitations 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Pericardial effusion 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Right ventricular failure 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Abdominal pain 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Ascites 3/52 (5.77%)
[secondary trial, Adverse Events cohort 2] Total: 25/49 (51.02%)
[secondary trial, Adverse Events cohort 2] Anaemia 2/49 (4.08%)
[secondary trial, Adverse Events cohort 2] Pancytopenia 0/49 (0.00%)
[secondary trial, Adverse Events cohort 2] Acute myocardial infarction 1/49 (2.04%)
[secondary trial, Adverse Events cohort 2] Atrial fibrillation 1/49 (2.04%)
[secondary trial, Adverse Events cohort 2] Cardiac failure 0/49 (0.00%)
[secondary trial, Adverse Events cohort 2] Cardiogenic shock 0/49 (0.00%)
[secondary trial, Adverse Events cohort 2] Palpitations 1/49 (2.04%)
[secondary trial, Adverse Events cohort 2] Pericardial effusion 4/49 (8.16%)
[secondary trial, Adverse Events cohort 2] Right ventricular failure 0/49 (0.00%)
[secondary trial, Adverse Events cohort 2] Abdominal pain 1/49 (2.04%)
[secondary trial, Adverse Events cohort 2] Ascites 0/49 (0.00%)
*** Answer:
Let's think step by step. Among all evidence related to the secondary trial,
there is no mention of breathing related issues.
Then this is 0 patent who experienced breathing realted issues in the secondary trial.
Therefore, answer is "0"

#########################################
*** Question: What is the proportion of patients experienced nausea in the secondary trial?
*** Evidence:
[primary trial, Adverse Events cohort 1] Total: 9/23 (39.13%)
[primary trial, Adverse Events cohort 1] Anemia 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Diarrhea 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Nausea 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Fracture 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] ALT 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] AST 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Glucose, high 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Limb Pain 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Ataxia 2/23 (8.70%)
[primary trial, Adverse Events cohort 1] Neurology - Other 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Seizure 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Syncope 1/23 (4.35%)
[primary trial, Adverse Events cohort 1] Confusion 1/23 (4.35%)
[secondary trial, Adverse Events cohort 1] Total: 14/52 (26.92%)
[secondary trial, Adverse Events cohort 1] Tachycardia * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Atrial fibrillation * 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Duodenal ulcer * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Gastric ulcer * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Nausea * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Abdominal pain * 0/52 (0.00%)
[secondary trial, Adverse Events cohort 1] Mucosal inflammation * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 1] Appendicitis * 1/52 (1.92%)
[secondary trial, Adverse Events cohort 2] Total: 4/20 (20.00%)
[secondary trial, Adverse Events cohort 2] Tachycardia * 0/20 (0.00%)
[secondary trial, Adverse Events cohort 2] Atrial fibrillation * 1/20 (5.00%)
[secondary trial, Adverse Events cohort 2]
```

```
    Duodenal ulcer * 0/20 (0.00%)
[secondary trial, Adverse Events cohort 2]
    Gastric ulcer * 0/20 (0.00%)
[secondary trial, Adverse Events cohort 2]
    Nausea * 0/20 (0.00%)
[secondary trial, Adverse Events cohort 2]
    Abdominal pain * 1/20 (5.00%)
[secondary trial, Adverse Events cohort 2]
    Mucosal inflammation * 0/20 (0.00%)
[secondary trial, Adverse Events cohort 2]
    Appendicitis * 0/20 (0.00%)
*** Answer:
Let’s think step by step. From evidence "[
    secondary trial, Adverse Events cohort 1]
    Nausea * 1/52 (1.92%)"
and "[secondary trial, Adverse Events cohort 2]
    Nausea * 0/20 (0.00%)",
number of patients experienced nausea in the
    secondary trial is 1 + 0 = 1,
and the total number of patients in the
    secondary trial is 52 + 20 = 72.
Then the proportion of patients experienced
    nausea in the secondary trial is 1/72.
Therefore, answer is "1/72"
```

C.13 QA examples: eligibility

The following is the full list of 3 examples used for custom fine-tuned QA module for `section ID = Eligibility`:

```
##########################################
*** Question: What is the minimum life
    expectancy for the primary trial?
*** Evidence:
[primary trial, Inclusion Criteria] Ambulatory,
    female patients with an age 18 years
[primary trial, Inclusion Criteria] Patients
    with ECOG Performance status of 2
[primary trial, Inclusion Criteria] Patient who
    have estimated life expectancy of more than
    six months
[primary trial, Inclusion Criteria] No evidences
     of hemorrhage
[primary trial, Exclusion Criteria] Male
    patients
[primary trial, Exclusion Criteria] Patients
    weighing <45 Kg
[primary trial, Exclusion Criteria] Patients
    with prior bone marrow or stem cell
    transplantation
[primary trial, Exclusion Criteria] Patients
    with seropositivity for HIV or HBV or HCV
[primary trial, Exclusion Criteria] Known cases
    of Sickle Cell Anemia
[primary trial, Exclusion Criteria] Pregnant and
     Breast feeding women.
*** Answer:
Let’s think step by step. From evidence "[
    primary trial, Inclusion Criteria] Patient
    who have estimated life expectancy of more
    than six months",
minimum life expectancy for the primary trial is
     6 months.
Therefore, answer is "6 months"

##########################################
*** Question: Are patients other than White or
    Asian excluded from the primary trial?
*** Evidence:
[primary trial, Inclusion Criteria] Diagnosis of
     breast cancer
[primary trial, Inclusion Criteria] Ability to
    use internet
[primary trial, Inclusion Criteria] Ability to
    read and understand Danish
[primary trial, Exclusion Criteria] Diagnosis of
     primary lymphedema
[primary trial, Exclusion Criteria] Metastatic
    or inflammatory breast cancer
[primary trial, Exclusion Criteria] Planned
    hospitalization or surgery within the next
    twelve weeks
[primary trial, Exclusion Criteria] Planned use
    of chemotherapy within the next 6 weeks
[secondary trial, Inclusion Criteria]
    Participants must be at least 21 years of
    age.
[secondary trial, Inclusion Criteria]
    Participants must not be pregnant.
[secondary trial, Inclusion Criteria]
    Participants can be from any racial or
    ethnic origin.
[secondary trial, Inclusion Criteria]
    Participants with in situ breast cancer are
    eligible.
[secondary trial, Inclusion Criteria]
    Participant must give informed consent.
[secondary trial, Exclusion Criteria] Patients
    with bilateral breast cancer are not
    eligible.
[secondary trial, Exclusion Criteria] Patients
    who have had previous radiation therapy to
    the breast or chest are not eligible.
[secondary trial, Exclusion Criteria] Patients
    cannot have had breast reconstructions,
    implants, and/or expanders.
*** Answer:
Let’s think step by step. Among all evidence
    related to the primary trial,
there is no mention of White, Asian, or racial
    criteria.
Therefore, answer is "No"

##########################################
*** Question: Are patients with HER2-positive
    breast cancer eligible for the secondary
    trial?
*** Evidence:
[primary trial, Inclusion Criteria] Women 20
    years
[primary trial, Inclusion Criteria]
    Histologically or cytologically confirmed
    diagnosis of adenocarcinoma originating in
    the breast
[primary trial, Exclusion Criteria] Number of
    prior chemotherapy lines of treatment in the
     metastatic setting 3[secondary trial,
    Inclusion Criteria] Adult participants >/=18
     years of age
[secondary trial, Inclusion Criteria] HER2-
    positive breast cancer
[secondary trial, Inclusion Criteria]
    Participants must have measurable and/or non
    -measurable disease which must be evaluable
    per Response Evaluation Criteria in Solid
    Tumors (RECIST) 1.1
[secondary trial, Inclusion Criteria] Eastern
    Cooperative Oncology Group (ECOG)
```

```
     Performance Status 0 or 1
[secondary trial, Inclusion Criteria] Adequate
     organ function as determined by laboratory
     results
[secondary trial, Exclusion Criteria] History of
      prior (or any) chemotherapy for metastatic
     breast cancer or recurrent locally advanced
     disease
[secondary trial, Exclusion Criteria] An
     interval of <6 months from the last dose of
     vinca-alkaloid or taxane cytotoxic
     chemotherapy until the time of metastatic
     diagnosis
[secondary trial, Exclusion Criteria] Hormone
     therapy <7 days prior to randomization
[secondary trial, Exclusion Criteria]
     Trastuzumab therapy and/or lapatinib (neo-
     or adjuvant setting) <21 days prior to
     randomization
[secondary trial, Exclusion Criteria] Prior
     trastuzumab emtansine or pertuzumab therapy
*** Answer:
Let's think step by step. From evidence "[
     secondary trial, Inclusion Criteria] HER2-
     positive breast cancer",
HER2-positive breast cancer patients are
     eligible for the secondary trial.
Therefore, answer is "Yes"
```

C.14 AG examples: results

The following is the full list of 3 examples used for custom fine-tuned AG module for `section ID = Results`:

```
##########################################
*** Statement: Cohort 1 of the primary trial had
      a longer median, maximum and minimum pfs
     than cohort 2.

[Pair 1]
*** Question: What is the median pfs of cohort 1
      in the primary trial?
*** Answer: 103 Days

[Pair 2]
*** Question: What is the maximum pfs of cohort
     1 in the primary trial?
*** Answer: 83 Days

[Pair 3]
*** Question: What is the minimum pfs of cohort
     1 in the primary trial?
*** Answer: 128 Days

[Pair 4]
*** Question: What is the median pfs of cohort 2
      in the primary trial?
*** Answer: 81 Days

[Pair 5]
*** Question: What is the maximum pfs of cohort
     2 in the primary trial?
*** Answer: 48 Days

[Pair 6]
*** Question: What is the minimum pfs of cohort
     2 in the primary trial?
*** Answer: 95 Days

*** Response: Let's think step by step.
From Pair 1 and Pair 4, for the median pfs, 103
     Days > 81 Days, so cohort 1 > cohort 2.
From Pair 2 and Pair 5, for the maximum pfs, 83
     Days > 48 Days, so cohort 1 > cohort 2.
From Pair 3 and Pair 6, for the minimum pfs, 128
      Days > 95 Days, so cohort 1 > cohort 2.
Therefore, the statement is true.
Agreement: True

##########################################
*** Statement: cohort 1 of the secondary trial
     had a much longer median PFS than cohort 1
     of the primary trial.

[Pair 1]
*** Question: What is the median PFS of cohort 1
      of the primary trial?
*** Answer: 89

*** Response: Let's think step by step.
median PFS for the cohort 1 of the secondary
     trial is not available in the given question
     -answer pairs.
Therefore, the statement is false.
Agreement: False

##########################################
*** Statement: In the primary trial, Dasatinib,
     70 mg, Twice Daily results in a better
     median PFS than Dasatinib, 100 mg, Daily.
     The opposite was true in the secondary trial
     .

[Pair 1]
*** Question: Which cohort in the primary trial
     uses Dasatinib, 70 mg, Twice Daily?
*** Answer: cohort 2

[Pair 2]
*** Question: Which cohort in the primary trial
     uses Dasatinib, 100 mg, Daily?
*** Answer: cohort 1

[Pair 3]
*** Question: What is the median PFS of cohort 1
      of the primary trial?
*** Answer: 10.3 weeks

[Pair 4]
*** Question: What is the median PFS of cohort 2
      of the primary trial?
*** Answer: 15.3 weeks

*** Response: Let's think step by step.
From Pair 1, In the primary trial, Dasatinib, 70
      mg, Twice Daily, is given to cohort 2.
From Pair 4, The median PFS of cohort 2 is 15.3
     weeks.
From Pair 2, In the primary trial, Dasatinib,
     100 mg, Daily, is given to cohort 1.
From Pair 3, The median PFS of cohort 1 is 10.3
     weeks.
So, in the primary trial, Dasatinib, 70 mg,
     Twice Daily results in a better median PFS
     than Dasatinib, 100 mg, Daily.
The cohort in the secondary trial uses Dasatinib,
      70 mg, Twice Daily is not available in the
     given question-answer pairs.
```

```
Therefore, the statement is false.
Agreement: False
```

C.15 AG examples: intervention

The following is the full list of 3 examples used for custom fine-tuned AG module for `section ID = Intervention`:

```
#########################################
*** Statement: Patients in cohort 1 of the
    primary trial may receive gradually
    increasing doses of Afatinib monotherapy, up
     to 160% of the starting dose.

[Pair 1]
*** Question: Do patients in cohort 1 of the
    primary trial receive gradually increasing
    doses of Afatinib monotherapy?
*** Answer: Yes

[Pair 2]
*** Question: What is the starting dose of
    Afatinib monotherapy for the cohort of the
    primary trial?
*** Answer: 40mg

[Pair 3]
*** Question: What is the maximum dose of
    Afatinib monotherapy for the cohort of the
    primary trial?
*** Answer: 50mg

*** Response: Let's think step by step.
From Pair 1, Patients in cohort 1 of the primary
     trial receive gradually increasing doses of
     Afatinib monotherapy.
From Pair 2 and Pair 3, the maximum dose of
    Afatinib monotherapy is 50mg and the
    starting dose of Afatinib monotherapy is 40
    mg,
so the dose of Afatinib monotherapy increases up
     to 50mg/40mg = 1.25 = 125% of the starting
    dose.
Therefore, the statement is false.
Agreement: False

#########################################
*** Statement: Between all cohorts in the
    primary trial and the secondary trial Omega
    -3-fatty Acids are only used in one cohort.

[Pair 1]
*** Question: Is Omega-3-fatty acids used in
    cohort 1 in the primary trial?
*** Answer: Yes

[Pair 2]
*** Question: Is Omega-3-fatty acids used in
    cohort 2 in the primary trial?
*** Answer: No

[Pair 3]
*** Question: Is Omega-3-fatty acids used in
    cohort 1 in the secondary trial?
*** Answer: No

[Pair 3]
*** Question: Is Omega-3-fatty acids used in
    cohort 2 in the secondary trial?
*** Answer: No

*** Response: Let's think step by step.
From Pair 1, Pair 2, Pair 3, and Pair 4, cohort
    1 in the primary trial is the only cohort
    that Omega-3-fatty Acids are used.
Therefore, the statement is true.
Agreement: True

#########################################
*** Statement: All Participants in the primary
    trial and the secondary trial are receiving
    the same daily dose of Degarelix injected
    into the muscle, for the same duration of
    time.

[Pair 1]
*** Question: What is the daily dose of
    Degarelix injected into the muscle for the
    participants in cohort 1 in the primary
    trial?
*** Answer: 240 mg

[Pair 2]
*** Question: What is the daily dose of
    Degarelix injected into the muscle for the
    participants in cohort 2 in the primary
    trial?
*** Answer: 240 mg

[Pair 3]
*** Question: What is the duration of time that
    Degarelix injected into the muscle for the
    participants in cohort 1 in the primary
    trial?
*** Answer: 28 days

*** Response: Let's think step by step.
From Pair 1, the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 1 in the primary
    trial is 240 mg.
From Pair 2, the daily dose of Degarelix
    injected into the muscle for the
    participants in cohort 2 in the primary
    trial is 240 mg.
the daily dose of Degarelix injected into the
    muscle for the participants in the secondary
     trial is not available in the given
    question-answer pairs.
Therefore, the statement is false.
Agreement: False
```

C.16 AG examples: adverse events

The following is the full list of 3 examples used for custom fine-tuned AG module for `section ID = Adverse Events`:

```
#########################################
*** Statement: The same number of cases of
    Neutropenia, Febrile neutropenia and
    Pancytopenia are observed in patients from
    cohort 1 of the primary trial.

[Pair 1]
```

```
*** Question: How many patients from cohort 1 of
     the primary trial had Neutropenia?
*** Answer: 3

[Pair 2]
*** Question: How many patients from cohort 1 of
     the primary trial had Febrile neutropenia?
*** Answer: 2

[Pair 3]
*** Question: How many patients from cohort 1 of
     the primary trial had Pancytopenia?
*** Answer: 2

*** Response: Let's think step by step.
From Pair 1, the number of cases of Neutropenia
    from cohort 1 of the primary trial is 3.
From Pair 2, the number of cases of Febrile
    neutropenia from cohort 1 of the primary
    trial is 2.
From Pair 3, the number of cases of Pancytopenia
     from cohort 1 of the primary trial is 2.
Therefore, the statement is false.
Agreement: False

#########################################
*** Statement: Between both of the patient
    cohort of the primary trial and the
    secondary trial there was only a single
    patient who suffered heart failure.

[Pair 1]
*** Question: How many patients suffered with
    heart failure in the primary trial?
*** Answer: 0

[Pair 2]
*** Question: How many patients suffered with
    heart failure in the secondary trial?
*** Answer: 1

*** Response: Let's think step by step.
From Pair 1 and Pair 2, there is only one
    patient who suffered with heart failure
    between both of the patient cohorts.
Therefore, the statement is true.
Agreement: True

#########################################
*** Statement: the primary trial had a lower
    total percentage of patients experiencing
    adverse events compared to the secondary
    trial.

[Pair 1]
*** Question: What is the percentage of patients
     experienced adverse events in the primary
    trial?
*** Answer: 9/51 (17.65%)

[Pair 2]
*** Question: What is the percentage of patients
     experienced adverse events in the secondary
     trial?
*** Answer: 5/38 (13.16%)

*** Response: Let's think step by step.
From Pair 1 and Pair 2, 17.65% > 13.16%, so
    primary trial > secondary trial in total
    percentage of patients experiencing adverse
    events.
Therefore, the statement is false.
Agreement: False
```

C.17 AG examples: eligibility

The following is the full list of 3 examples used for custom fine-tuned AG module for `section ID = Eligibility`:

```
#########################################
*** Statement: Patients must have a life
    expectancy over a year to participate in the
     primary trial.

[Pair 1]
*** Question: What is the minimum life
    expectancy for the primary trial?
*** Answer: 6 months

*** Response: Let's think step by step.
From Pair 1, patients with life expectancy over
    6 months can participate in the primary
    trial.
Therefore, the statement is false.
Agreement: False

#########################################
*** Statement: Agatha had her 53rd birthday last
     week, she has a histologically confirmed
    adenocarcinoma of the breast, with no
    evidence of metastatic disease. Agatha is of
     white british origin. she is eligible for
    the primary trial but not the secondary
    trial, due to her age.

[Pair 1]
*** Question: Are women of 53 years old eligible
     for the primary trial?
*** Answer: Yes

[Pair 2]
*** Question: Are women of 53 years old eligible
     for the secondary trial?
*** Answer: Yes

[Pair 3]
*** Question: Are patients with a histologically
     confirmed adenocarcinoma of the breast with
     no evidence of metastatic disease eligible
    for the primary trial?
*** Answer: Yes

[Pair 4]
*** Question: Are patients of white british
    origin eligible for the primary trial?
*** Answer: Yes

*** Response: Let's think step by step.
From Pair 1 and Pair 2, a women of 53 years old
    is eligible for the both trials.
From Pair 3, patients with a histologically
    confirmed adenocarcinoma of the breast with
    no evidence of metastatic disease are
    eligible for the primary trial.
From Pair 4, patients of white british origin
    are eligible for the primary trial.
Therefore, the statement is true.
Agreement: True
```

```
########################################
*** Statement: All cancer stages are accepted
    for the primary trial and the secondary
    trial.

[Pair 1]
*** Question: Are patients with any cancer stage
     eligible for the primary trial?
*** Answer: Yes

[Pair 2]
*** Question: Are patients with any cancer stage
     eligible for the secondary trial?
*** Answer: No

*** Response: Let’s think step by step.
From Pair 1, patients with any cancer stage are
    eligible for the primary trial.
From Pair 2, not all patients with any cancer
    stage are eligible for the primary trial.
Therefore, the statement is false.
Agreement: False
```