SweSAT-1.0: The Swedish University Entrance Exam as a Benchmark for Large Language Models

Murathan Kurfalı[1] **Shorouq Zahra**[1] **Evangelia Gogoulou**[1]
Luise Dürlich[1,2] **Fredrik Carlsson**[1] **Joakim Nivre**[1,2]
[1]RISE Research Institutes of Sweden, Sweden
[2]Uppsala University, Sweden

Abstract

This introduces SweSAT-1.0, a new benchmark dataset created from the Swedish university entrance exam (Högskoleprovet) to assess large language models in Swedish. The current version of the benchmark includes 867 questions across six different tasks, including reading comprehension, mathematical problem solving, and logical reasoning. We find that some widely used open-source and commercial models excel in verbal tasks, but we also see that all models, even the commercial ones, struggle with reasoning tasks in Swedish. We hope that SweSAT-1.0 will facilitate research on large language models for Swedish by enriching the breadth of available tasks, offering a challenging evaluation benchmark that is free from any translation biases.

1 Introduction

The recent progress in language modeling has significantly expanded the generalization capabilities of large language models (LLMs). Models such as Llama 3.1 (Dubey et al., 2024), Gemma (Team et al., 2024), and GPT-4 (Achiam et al., 2023) have demonstrated remarkable performance across a wide range of NLP tasks, exceeding the expectations researchers held just a few years ago. Consequently, many existing benchmarks are found to be inadequate due to their task-specific nature, focusing narrowly on traditional classification problems and failing to capture the full spectrum of language understanding capabilities of modern LLMs. Benchmarks such as SuperGLUE (Wang et al., 2019) and XTREME (Hu et al., 2020) predominantly assess specific NLP tasks, limiting their ability to evaluate the broader, more generalized language capabilities that contemporary LLMs are seemingly capable of.

This issue is even more crucial for languages other than English, which are often evaluated on translated benchmarks that are prone to numerous biases and quality issues. To address this gap, we follow the tradition of using standardized exams (Hendrycks et al., 2020; Achiam et al., 2023) and introduce the first version of a Swedish benchmark called SweSAT-1.0[1] sourced from the Swedish university entrance exam, *Swedish Scholastic Aptitude Test* ('Högskoleprovet' in Swedish). The exam encompasses both verbal and quantitative reasoning tests across several sub-categories, such as reading comprehension, mathematical problem solving, and logical reasoning.

SweSAT-1.0 is sourced from the last eight exams over the past five years. The benchmark is prepared through automatic parsing of the exam files, followed by manual checks to correct any parsing errors. It currently comprises 867 questions and has the following advantages over most existing benchmarks: i) it is free from translation biases and culturally irrelevant content; ii) it allows researchers to control for data contamination[2] as the exact administration dates of the exams are known; iii) it broadens the range of tasks available for evaluation in Swedish; and iv) it indirectly allows a comparison against the real exam takers as the results are publicly available.

In addition to presenting the benchmark, we evaluate a wide range of popular multilingual and Swedish-oriented LLMs. The results show that while multilingual LLMs outperform their Swedish-oriented counterparts, even the commercial models fail at solving the reasoning tasks in Swedish, highlighting a crucial shortcoming of the existing LLMs. We hope that the benchmark will

[1]The dataset can be accessed here: `https://github.com/NLP-RISE/swesat`

[2]Data contamination occurs when some or all of the test data is inadvertently included in the training set (Li et al., 2024).

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 331–339
March 3-4, 2025

Section	Total # questions	Description
ORD	160	*Vocabulary*: Tests the understanding of in-domain words and synonyms.
LÄS	160	*Reading comprehension:* Assesses the ability to make inference from a text.
MEK	160	*Sentence completion*: Assesses the ability to complete sentences via cloze tests.
XYZ	157	*Mathematical problem-solving*: Tests arithmetic, algebra, geometry, statistics, and functions.
KVA	140	*Quantitative comparisons*: Measures the ability to compare quantities in math concepts.
NOG	90	*Data sufficiency*: Evaluates the ability to determine if data is sufficient for solving a problem.

Table 1: Overview of exam sections in SweSAT-1.0, with total number of questions per section.

contribute to the evaluation of LLM performance in Swedish and encourage further research and development in multilingual contexts.

2 Related Work

There are a few benchmarks specifically designed to evaluate NLP models in Swedish, with SuperLim (Berdičevskis et al., 2023) and ScandEval (Nielsen, 2023) being the most prominent examples. Created as the Swedish counterpart to SuperGLUE, SuperLim is a comprehensive test suite that consists of 15 tasks, such as word analogy, pronoun resolution, and text summarization.[3] If not adapted from English through translation, the featured datasets are either constructed by reformatting pre-existing tasks or created from scratch using pre-existing corpora. The reliance on pre-existing datasets raises concerns about data contamination, and the use of translation could introduce bias, which signals the need for new and complementary evaluation datasets. ScandEval (Nielsen, 2023; Nielsen et al., 2024), on the other hand, provides a multilingual evaluation suite spanning a subset of North Germanic languages, among them Swedish. Despite broad task coverage, the majority of ScandEval datasets are revisited versions of existing datasets, which again raises concerns about whether data contamination and the use of machine translation could undermine the evaluation process.

3 Dataset Description

SweSAT-1.0 is a benchmark dataset sourced from the publicly available *Swedish Scholastic Aptitude Test*,[4] a standardized Swedish university entrance exam. The exam is written and administered by the Swedish Council for Higher Education and used for admission to higher education in Sweden. The exam consists of two main parts: verbal and quantitative, each containing four sections. Each exam includes 160 multiple-choice questions taken over a single day, lasting almost 8 hours (including breaks). This exam has been selected for its high quality; since it is written specifically to assess students' verbal and quantitative reasoning skills in Swedish, we eliminate the risk of cultural and linguistic biases.

Sample questions can be found in Appendix C. We refer interested readers to Stage and Ögren (2004) for more detailed information on the exam.

3.1 Dataset Construction

The dataset was constructed through a semi-automatic process. Although the exam files are available in PDF format, extracting the content correctly proved challenging due the documents' structure and formatting. For the verbal part, we employed pdfplumber,[5] a popular Python library for PDF parsing. This approach worked well for extracting plain text but struggled with recognizing and preserving the format of mathematical expressions in the quantitative sections. Therefore, we adopted a different method for quantitative questions: we first converted each page into a high-resolution image, then performed OCR using GPT-4o (2024-08-06) with a detailed prompt (see Appendix A) to accurately capture both the text and mathematical formulas. The latter were represented in LaTeX in a consistent format, following common practice (Wang et al., 2023; Zhang et al., 2023). Despite our best efforts, we discovered that there were various errors in the final output, such as improper handling of hyphenated words at line breaks, italicized words jumping onto the wrong lines, or LaTeX formatting issues. Therefore, each exam was manually checked and corrected for errors to ensure accuracy and consistency.

[3] We note that one word-level task in SuperLim is directly taken from the *ORD* section of SweSAT (see Table 1).

[4] https://www.studera.nu/hogskoleprov

[5] https://github.com/jsvine/pdfplumber

Model	ORD	LAS	MEK	XYZ	KVA	NOG	Average
Aya-23-8B	43.12	40.00	40.94	18.75	18.75	10.42	28.66
Gemma-2-9b	85.62	82.50	86.25	31.77	30.31	31.77	58.04
Gemma-2-27b	91.56	90.62	90.94	37.50	36.25	32.29	63.19
GPT-SW3-1.3b	16.88	22.50	25.94	18.23	21.25	9.38	19.03
GPT-SW3-6.7b-v2	20.00	21.25	25.62	17.19	19.38	11.46	19.15
GPT-SW3-20b	21.56	30.63	30.31	18.75	22.50	12.50	22.71
AI-Sweden/Llama-3-8B	71.25	56.25	59.69	21.88	20.31	13.02	40.40
Llama-3-8B	68.44	65.00	55.62	18.75	26.25	25.00	43.18
Llama-3.1-8B	80.31	69.38	58.75	20.83	31.25	18.23	46.46
GPT-4o-mini (2024-07-18)	97.50	84.38	96.25	32.29	38.12	35.42	63.99
GPT-4o (2024-08-06)	100.0	92.50	99.38	47.40	45.62	45.83	71.79

Table 2: Average performance of baseline models across question types on the entire SweSat 1.0.

3.2 The SweSAT-1.0 Dataset

SweSAT-1.0 includes the last five years of the exam (from 2020 to 2024) held over eight different sessions.[6] Following our primary focus on evaluating text-based language models in Swedish, SweSAT-1.0 includes only the verbal and quantitative reasoning sections that do not require multimodal inputs, thus omitting the entire section of DTK (Diagrams, Tables, and Maps) as well as any question that requires visual information to solve. The ELF (English Reading Comprehension) section is also excluded from the dataset since our primary focus is on Swedish.[7] The dataset currently comprises 867 questions, covering six question types, as shown in Table 1. All questions are in the multiple-choice format: ORD and NOG sections have five options whereas the remaining sections have only four.

Alongside the questions, we prompt the models using the official exam instructions to simulate the real exam-taking scenario. The original instructions include an explanation of the exam section, and one sample question and answer for five of the sections included in this dataset: ORD, MEK, KVA, NOG, and XYZ. However, the sample question and its answer in the XYZ section is excluded as it contains figures incompatible with the benchmark setup.

This results in a mix of one-shot and zero-shot prompts. We exclude the sample questions and answers in order to conduct the experiments using a zero-shot version of the exam instructions. The zero-shot version of these instructions (which excludes any example questions) as well as the mixed one-shot version are both included in the dataset release to facilitate a standardized evaluation across all sections.

4 Baselines

In this section, we evaluate the performance of a range of LLMs, each with different levels of Swedish coverage during training, on SweSAT-1.0. The primary purpose of this baseline evaluation is to evaluate the dataset itself by analyzing how some of the most popular LLMs perform on its tasks to ensure that the dataset is sufficiently challenging and valuable as a benchmark. By doing so, we also provide reference scores for future studies while exploring the current capabilities of LLMs in Swedish. Our evaluation includes a range of instruction-tuned open-source models such as GPT-SW3 (Ekgren et al., 2024), Gemma-2 (Team et al., 2024), Aya (Üstün et al., 2024), Llama 3 and 3.1 (Dubey et al., 2024), as well as the commercial GPT-4o-mini (2024-08-06) and GPT-4o (2024-07-18) models (Achiam et al., 2023). The entire model list can be found in Appendix B.

4.1 Experimental Setup

We use the original exam instructions (excluding the sample questions) as zero-shot prompts to assess the models' performance under authentic exam-taking conditions. To ensure adherence to these instructions, we add a brief directive[8] at

[6] At the time when the dataset was constructed, the 2024 fall exam has not yet been held, and only the spring exam is available for 2020.

[7] Note that the ELF sections are not publicly available, so these questions could not be included in the benchmark.

[8] *Svara endast med bokstaven på det rätta alternativet utan någon förklaring* ('Answer only with the letter of the correct option without any explanation').

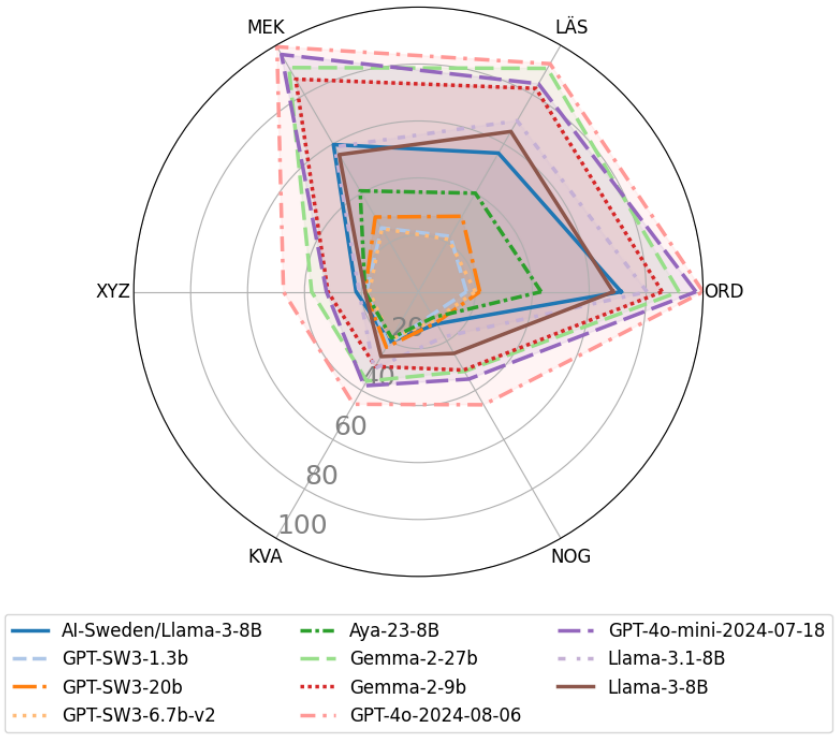


Figure 1: Radar chart comparing model performances across different question types

the end of each question, prompting the models to return a single letter as the desired response format. The models may at most output a single token; any output that does not exactly match one of the allowed answer letters (A, B, C, D, or E; depending on the question type) is discarded. Questions are presented one by one, with reading passages repeated in the prompt before each LÄS question. Answers are generated using greedy decoding[9] to ensure a deterministic output, hence reproducibility, by selecting the highest probability response at each step. As all questions are multiple-choice, we use accuracy as our evaluation metric.

4.2 Results

The average performance of baseline models across eight exams is shown in Table 2. All models perform markedly better on the verbal sections (MEK, LÄS, and ORD), with Gemma models achieving around 90% accuracy and GPT-4o achieving almost a perfect score in the MEK and ORD sections. On the other hand, quantitative sections yield significantly lower scores, with even GPT-4o failing on the majority of questions. Swedish-oriented models — all models in the GPT-SW3 family in addition to a fine-tuned Llama 3 version — consistently show lower accuracy across all question types. To note a special case, we find that the aforementioned Llama 3 instruct-variant, fine-tuned on The Nordic Pile (Öhman et al., 2023),[10] exhibits better performance than all evaluated GPT-SW3 models. Yet, it achieves slightly lower average accuracy than the original Llama 3 on five of the eight exams. This raises questions on whether continued pre-training on a mix of Scandinavian languages is useful for this task, or whether it may depend on the nature of the selected dataset.

The differences among models across question types are further illustrated in Figure 1. The results suggest that current LLMs have significant limitations in quantitative reasoning tasks in Swedish. Furthermore, we also analyze the patterns in the way models provide answers through confusion matrices (see Appendix D). GPT-SW3 models are observed to frequently select the same options (e.g., consistently choosing A or alternating between A and D in the case of the 20B version), which highlights potential shortcomings in following instructions. However, the selected options for other models are more evenly distributed across the potential answers, suggesting better task understanding,

[9] For GPT-4o models, we set the temperature to $1e^{-9}$.

[10] A dataset comprised of a mix of Scandinavian languages and English.

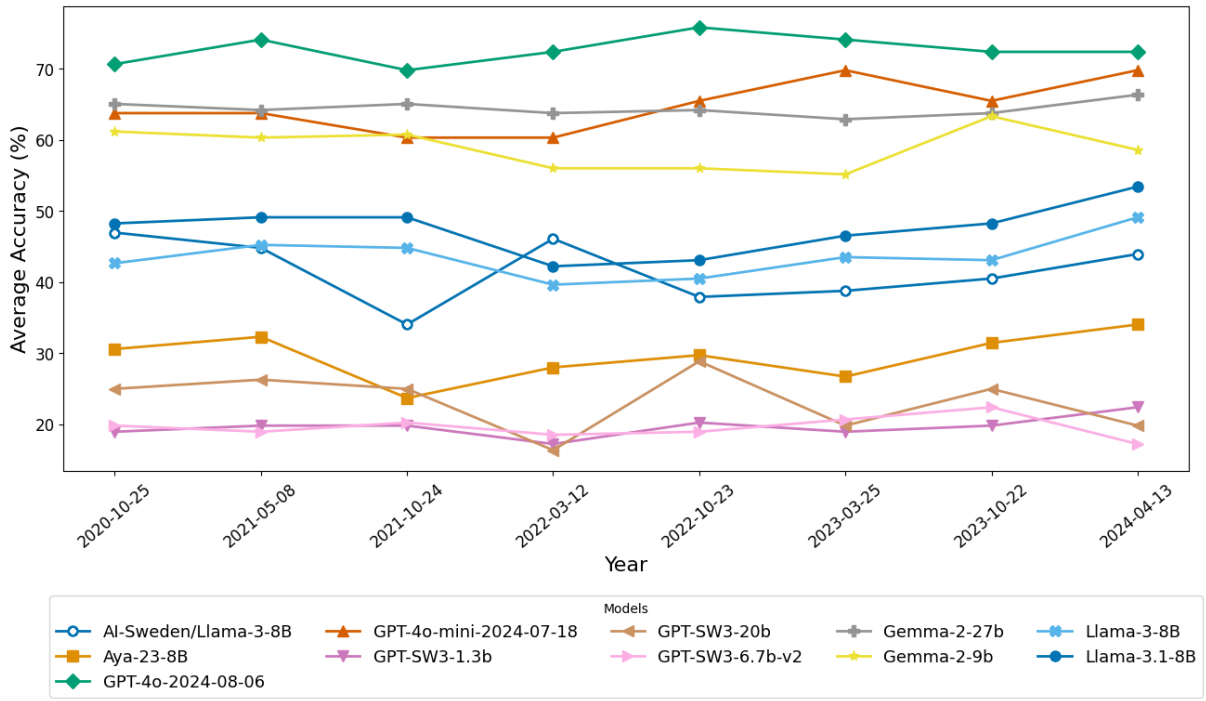


Figure 2: Average performance of baseline models across years

even though the correct option is not consistently identified. Yet, it should also be noted that this evaluation represents a particularly challenging setting where we require the models to produce only the correct answer without using techniques like chain-of-thought prompting (Wei et al., 2022) and without model-specific prompt engineering.

Finally, we investigate the potential impact of data contamination on LLM performance. As shown in Figure 2, all models exhibit a highly consistent performance across exam years, with an average standard deviation of only 2.8% in accuracy. This suggests that the contamination effect is absent and that the exam difficulty is consistent across years.

5 Conclusion

In this paper, we present a comprehensive benchmark to evaluate LLMs' various abilities in Swedish, using the university entrance exam. We believe our benchmark provides a consistent framework for testing LLM performance across a range of tasks detailed above, with an option to control for data contamination in model training through exam timestamps. Our baseline evaluations reveal the high accuracy of multilingual models across verbal tasks compared to their Swedish-centric counterparts – but also the overall weakness of all tested models on the reasoning tasks.

Acknowledgments

We gratefully acknowledge the support of the Swedish Research Council (grant no. 2022-02909). The experiments with the open-source LLMs were enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 202206725 and 2024-01506. We thank NAISS for providing computational resources under Project 2024/22-211. We also thank the reviewers for their valuable feedback and suggestions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT–4 technical report. *arXiv preprint arXiv:2303.08774*.

Aleksandrs Berdičevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, et al. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalization. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2024. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. *arXiv preprint arXiv:2406.13469*.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The Nordic Pile: A 1.2tb Nordic dataset for language modeling. *arXiv preprint arXiv:2303.17183*.

Christina Stage and Gunilla Ögren. 2004. *The Swedish Scholastic Assessment Test (SweSAT): Development, Results, and Experiences*. EM nr 49. Umeå University, Department of Educational Measurement, Umeå, Sweden.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems 36*, pages 5484–5505.

A Parsing the Quantitative Questions

Parsing the quantitative part of the exams, i.e., sections containing mathematical expressions and visual elements, proved to be very challenging for standard PDF parsing libraries in accurately recovering complex mathematical equations. Therefore, we performed OCR on each page separately using GPT-4o with the following prompt:

The image contains an exam sheet with both text-based and visual-based questions. Your task is to extract only the text-based questions and answers and format them into the following JSON structure. If a question contains any actual visual (such as a diagram, shape, figure, graph, or table visible in the image), set `"is_accompanied_with_visual"` to `"yes"` and specify the type of the visual in the `"visual_type"` field (e.g., "diagram", "graph", etc.). In this case, set the `"question"` field to `"Visual required to solve this question"` and leave the `"answers"` field blank. If the question describes geometrical objects (like lines, points, or coordinates) but does not include an actual visible diagram, treat it as a text-only question. Set `"is_accompanied_with_visual"` to `"no"` and fully extract the question and answers, preserving all formulas and numbers exactly as shown.

JSON Format:

```
[
  {
    "question_number": <number>,
    "question": "<question_text_with_formulas>",
    "answers": {
      "a": "<option a>",
      "b": "<option b>",
      "c": "<option c>",
      "d": "<option d>"
    },
    "is_accompanied_with_visual": "<yes/no>",
    "visual_type": "<visual_type>",
    "question_type": "<XYZ/KVA/NOG/DTK>"
  }
]
```

Further Instructions:

- Fully extract text-based questions and answers exactly as shown without modification.
- Do not simplify or paraphrase any part of the question. Classify the question as XYZ, KVA, NOG, or DTK.
- All the math formulas must be represented as LaTeX code, surrounded by $ (e.g., $\frac{1}{3}$). Convert special math notations, such as $\surd$, into the corresponding LaTeX format. Wrap all the formulas with $ symbols.
- Pay extra attention to capturing exponents correctly. Be aware that there may be fractional exponents, such as `\( x^{\frac{5}{15}} \)`
- Distinguish clearly between similar characters, particularly "2" and "5" and "6" and "8", to avoid confusion.
- Pay close attention to capturing nested exponents and grouping symbols accurately. When encoding expressions, make sure to wrap exponents and nested exponents within braces {} to maintain the correct mathematical hierarchy. For example, `$\left( x^7 \right)^{\frac{1}{2}}$`.
- Validate the resulting LaTeX expression by ensuring it visually matches the intended structure of the original mathematical notation.
- Pay special attention to minus signs. Ensure that all minus signs are correctly included and accurately placed.
- Always encode expressions properly in LaTeX. Make sure to use '\' for LaTeX commands and wrap **all formulas** with $ symbols.
- Ensure all LaTeX functions, such as \times and \text, are used only within math mode (i.e., surrounded by $...$).

B Baseline models

Table 3 provides the repository names of the baselines models on `https://huggingface.co/`, alongside their simplified names used throughout the text. As for the OpenAI models, we used the (2024-07-18) release of GPT-4o-mini and the (2024-08-06) release of GPT-4o.

Simplified Name	HuggingFace model repository
Aya-23-8B	CohereForAI/aya-23-8B
Gemma-2-27b	google/gemma-2-27b-it
Gemma-2-9b	google/gemma-2-9b-it
GPT-SW3-1.3b	AI-Sweden-Models/gpt-sw3-1.3b-instruct
GPT-SW3-20b	AI-Sweden-Models/gpt-sw3-20b-instruct
GPT-SW3-6.7b-v2	AI-Sweden-Models/gpt-sw3-6.7b-v2-instruct
AI-Sweden/Llama-3-8B	AI-Sweden-Models/Llama-3-8B-instruct
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Llama-3-8B	meta-llama/Meta-Llama-3-8B-Instruct

Table 3: HuggingFace model repository names of the baseline models

C Example Questions

Figures 3 and 4 show sample questions from the NOG and KVA question types (respectively), as shown in the exam sheet.

D Confusion Matrices

Figure 5 presents confusion matrices summarizing model predictions over the entire SweSAT-1.0 dataset.

28. **Vad är medelvärdet av *a* och *b*?**

(1) Medelvärdet av (a + 5) och (b + 9) är lika med 10,5.

(2) Medelvärdet av a, (b − 1) och 3 är lika med 3.

Tillräcklig information för lösningen erhålls

A i (1) men ej i (2)
B i (2) men ej i (1)
C i (1) tillsammans med (2)
D i (1) och (2) var för sig
E ej genom de båda påståendena

What is the mean of a and b?

1. The mean of $(a + 5)$ and $(b + 9)$ is equal to 10.5.
2. The mean of a, $(b - 1)$ and 3 is equal to 3.

Sufficient information for solving the problem is obtained:

A from (1) but not from (2)
B from (2) but not from (1)
C from (1) and (2) together
D from both (1) and (2) each by itself
E not from the two statements

Figure 3: A sample from the NOG question type in Swedish (top) and translated to English (bottom)

13. Medelvärdet av de tre talen *x*, *y* och *z* är 12. Summan av *y* och *z* är 30.

Kvantitet I: *x*

Kvantitet II: 9

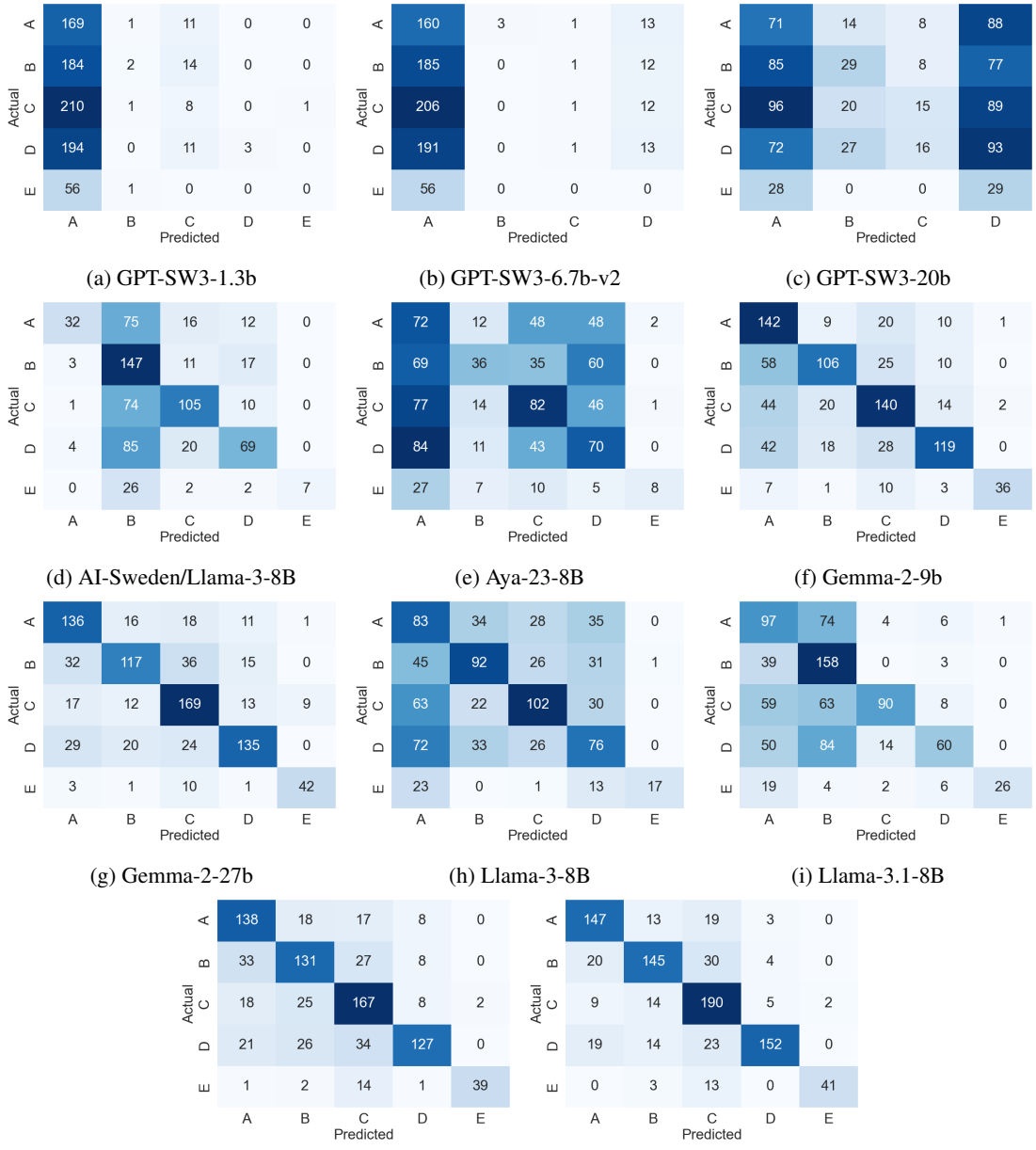
A I är större än II
B II är större än I
C I är lika med II
D informationen är otillräcklig

The mean value of the three numbers x, y and z is 12. The sum of y and z is 30.

1. Quantity I: x
2. Quantity II: 9

A I is greater than II
B II is greater than I
C I is equal to II
D The information is insufficient

Figure 4: A sample from the KVA question type in Swedish (top) and translated to English (bottom)



(a) GPT-SW3-1.3b (b) GPT-SW3-6.7b-v2 (c) GPT-SW3-20b

(d) AI-Sweden/Llama-3-8B (e) Aya-23-8B (f) Gemma-2-9b

(g) Gemma-2-27b (h) Llama-3-8B (i) Llama-3.1-8B

(j) GPT-4o-mini-2024-07-18 (k) GPT-4o-2024-08-06

Figure 5: Confusion matrices for the baseline models on SweSAT 1.0. Note that only two sections feature five options, hence the lower frequency of option E.