SweSpecQA: A Swedish Benchmark for Specialized Document-Native Question Answering

Murathan Kurfalı
RISE Research Institutes of Sweden, Sweden
Department of Linguistics and Philology, Uppsala University, Sweden
murathan.kurfali@lingfil.uu.se

Abstract

We introduce *SweSpecQA*, to our knowledge, the first Swedish benchmark for specialized institutional and professional QA with paired text and full-page PDF evaluation. It contains 1,500 questions from public Swedish examination and certification materials across six domains: EU aptitude, boating, antibiotic stewardship, interpreter licensing, physician licensing, and pharmacist licensing. Of these, 234 questions require an original visual or layout-dependent source element; the remaining 1,266 are text-eligible. All benchmark questions are drawn directly from Swedish-language source materials, avoiding translation artifacts in multilingual adaptations of English benchmarks. We evaluate 17 models under a closed-book protocol, reporting accuracy by domain, exam subtype, visual/layout dependence, modality, and prompting strategy, as well as whether models follow the requested answer format. On the text-eligible subset, every multimodal model performs worse with full source-page input than extracted text; PDF penalties range from 4.0 to 38.2 points under direct prompting and from 1.7 to 36.4 points under chain-of-thought prompting. Reasoning prompts often improve PDF performance for several open-weight VLMs, while sometimes reducing clean-text accuracy. These results show that document-native evaluation exposes failures hidden by text-only evaluation and that preprocessing remains important for reliable Swedish exam-style QA.

1 Introduction

Recent developments in large language models (LLMs) have led to their deployment in domains where performance depends not only on general language ability but also on specialized expertise and country-specific rules, such as medicine, law, government services, and professional examinations (Hendrycks et al., 2021; Wang et al., 2024). Evaluating these systems in the languages and institutions where they are used is therefore a practical necessity, yet most influential benchmarks are developed in English and adapted to other languages through translation (Hendrycks et al., 2021). Translation removes local institutional context, introduces fluency artifacts, and has been shown to inflate apparent model competence on the target language (Son et al., 2025; Huang et al., 2023).

Swedish evaluation resources have not kept up with the shift toward generative, instruction-following models. Existing suites such as SuperLim (Berdicevskis et al., 2023) and ScandEval (Nielsen, 2023) cover a range of NLU tasks but were designed primarily for discriminative evaluation of pretrained encoders and rely substantially on adapted or translated data. SweSAT-1.0 (Kurfalı et al., 2025) and the Swedish Medical LLM Benchmark (Moëll et al., 2025) are notable exceptions, providing evaluation for academic aptitude and clinical medicine respectively, but applied institutional knowledge across Swedish professional domains remains uncovered.

We introduce *SweSpecQA*, a translation-free benchmark for specialized Swedish question answering. To our knowledge, SweSpecQA is the first Swedish benchmark for specialized institutional and professional QA with paired text and full-page PDF evaluation. The benchmark includes questions that depend on figures, clinical images, navigation symbols, tables, charts, diagrams, or other page-layout information. It contains 1,500 questions collected directly from Swedish examination and certification materials across six institutional domains: medical exams, pharmacy exams, antibiotic-use guidelines, interpreter certification, EU aptitude/reasoning, and inshore boating certification. All questions were written for Swedish examinees and therefore preserve Swedish terminology, regulatory references, and institutional conventions that translated benchmarks cannot provide.

A central feature of *SweSpecQA* is its paired eval-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1998–2022
November 6-10, 2026 ©2026 Association for Computational Linguistics

uation design. In the **text track**, models receive parsed question text and answer alternatives. In the **PDF track**, models receive the full rendered source page or pages associated with the target question and only its visible question number, requiring them to locate the target question in the document, read it from the original layout, and solve it. We refer to the latter as *document-native* question answering because the model must answer directly from the original document page rather than from pre-extracted question text. Unlike benchmarks that crop the question image or separately supply the question text (Yue et al., 2024, 2025), we provide the full mapped source page or pages without any additional question text. This design lets us compare answering a question after extraction with answering it directly from the original document page. We evaluate models zero-shot using direct and chain-of-thought prompts, giving four settings: `text-direct`, `text-cot`, `pdf-direct`, and `pdf-cot`. On the matched text-eligible subset, every evaluated multimodal model performs worse in PDF mode, and the paired confidence intervals remain below zero for both prompting styles. The results therefore show a systematic performance drop when models answer questions from the original document pages, rather than simply reflecting that visual questions are harder than text-only questions. Our contributions are as follows:

- We introduce *SweSpecQA*, a 1,500-question Swedish-origin benchmark spanning six applied institutional and professional domains and 11 exam subtypes, with annotations for question type and visual/layout dependence.
- We provide paired text and full-page PDF evaluation tracks, enabling direct measurement of the gap between clean-text QA and document-native exam settings that require understanding the original layout, locating the question, and identifying the answer choices.
- We evaluate 17 OpenAI and open-weight LLMs under direct and reasoning prompts, and analyze the text–PDF performance gap, visual/layout-dependent questions, how CoT affects text and PDF performance differently, and whether models follow the requested answer format.

2 Related Work

Swedish and Scandinavian benchmarks. SuperLim (Berdicevskis et al., 2023) and ScandEval (Nielsen, 2023) are important resources for evaluating Swedish and Scandinavian language technology. They cover standard NLP and language-understanding tasks, including natural language inference, named entity recognition, and text classification, and rely substantially on adapted or translated source material. SweSAT-1.0 (Kurfalı et al., 2025) is a closer benchmark to our setting. It builds a translation-free benchmark from the Swedish university entrance examination and evaluates LLMs on verbal, quantitative, and logical questions. *SweSpecQA* follows the same motivation of using Swedish-origin exam material, but targets applied institutional and professional knowledge rather than academic aptitude.

Medical evaluation in Swedish has also started to receive attention. SMLB (Moëll et al., 2025) combines Swedish medical examination data with translated medical QA material. Its Medical Doctors component draws from the same Umeå University physician proficiency-exam source as our UMU Läkare SE subset, but from different exam years, while its PubMedQA component is translated. For the remaining SMLB components, we could not identify links to the original exam PDFs in the released benchmark materials; they therefore cannot support the paired extracted-text/full-page evaluation used here.

Language-oriented benchmarks. Several recent benchmarks use locally sourced examination material to evaluate models beyond English. C-Eval (Huang et al., 2023) uses questions from Chinese examinations across many subjects, while IndoMMLU (Koto et al., 2023) and ArabicMMLU (Koto et al., 2024) draw on Indonesian and Arabic examination material. KMMLU (Son et al., 2025) similarly collects expert-level questions from Korean examinations, and KMMLU-Pro (Hong et al., 2025) extends this setting with harder questions. These benchmarks motivate evaluation on questions written for the target language and educational or institutional context rather than translated from English.

Document and multimodal understanding. A second line of work evaluates models on documents and visual inputs. DocVQA (Mathew et al., 2021) and ChartQA (Masry et al., 2022) show that ques-

tion answering over documents and charts requires more than plain text understanding. MMMU (Yue et al., 2024) extends multimodal evaluation to college-level questions across several academic disciplines, and MMMU-Pro (Yue et al., 2025) studies harder settings where models must rely more directly on visual input.

SweSpecQA's PDF track is related to this work, but differs in one important way. We do not provide a cropped image of the target question or separately include the question text. Instead, the model receives the full mapped source page or pages and the visible question number. It must first find the question in this page input, then read and answer it. This makes the PDF track closer to the original exam setting and lets us measure errors from OCR, layout understanding, and question localization in addition to domain knowledge.

3 Benchmark Construction

This section describes the source materials, parsing pipeline, question formats, and evaluation tracks used to construct *SweSpecQA*. The source documents are available from their original providers. We release scripts for downloading and preprocessing these materials, constructing the benchmark, running the model evaluations, and reproducing the reported analyses.[1]

3.1 Data Sources

SweSpecQA draws from six publicly accessible Swedish examination and certification sources. Each source was selected on three criteria: the material originates in Swedish, an official answer key is available, and the exam targets applied institutional or professional competence rather than general academic knowledge. Across the six sources, the benchmark contains 11 subtypes: three EU reasoning subtypes, one Förarintyg subtype, one Strama subtype, two Tolkprov subtypes, two UMU Läkare subtypes, and two UU Apotekare subtypes. Table 1 summarizes the composition of each dataset. Detailed source descriptions, selection rules, and exclusions are provided in Appendix A. In brief:

EU MCQ SE. Swedish UHR practice material for EPSO verbal, numerical, and abstract reasoning.

Dataset	# Questions	%
EU MCQ SE	113	7.5%
Förarintyg SE	57	3.8%
Strama SE	200	13.3%
Tolkprov SE	120	8.0%
UMU Läkare SE	505	33.7%
UU Apotekare SE	505	33.7%
Total	1500	100.0%

Table 1: Dataset size by exam type in our benchmark.

Förarintyg SE. NFB practice material for the Swedish inshore boating certificate, covering navigation, safety, maritime rules, and chart interpretation; open-answer items and a near-duplicate Westcoast set are excluded.

Strama SE. Educational quizzes for Swedish healthcare professionals on antibiotic treatment guidelines, prescribing, resistance, and infection epidemiology.

Tolkprov SE. Kammarkollegiet practice material for authorized interpreters; we include the fixed-answer realia and Swedish-language proficiency components and exclude language-pair-specific terminology translation.

UMU Läkare SE. Umeå University theoretical proficiency examinations for physicians trained outside the EU/EEA; we include the MCQ and clinical-case components and exclude the English scientific-article component.

UU Apotekare SE. Uppsala University theoretical proficiency examinations for pharmacists trained outside the EU/EEA; we include Parts A and B and exclude the English scientific-article component.

3.2 Parsing and Item Extraction

The source material for *SweSpecQA* comes primarily from PDF documents. Because the sources differ substantially in layout, we use exam-specific methods to extract the questions and answer choices. For *Strama* and *Tolkprov*, the relatively regular layouts allow us to use PyMuPDF[2] together with source-specific parsing rules.

The other four sources contain more complex page layouts, including superscripts, subscripts, tables, figures, and mathematical notation. For these sources, we use GPT-5.4 to perform OCR-like transcription of the PDF pages into the common JSON

[1] https://github.com/MurathanKurfali/swespecqa

[2] https://pymupdf.readthedocs.io/

Model	Text				PDF			
	Direct		CoT		Direct		CoT	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Open-weight models								
Gemma-4-31B	89.7	89.5	91.2	89.9	66.7	51.4	76.8	68.3
Qwen3-VL-30B-A3B	79.0	75.7	70.9	69.0	39.6	34.3	60.2	49.8
InternVL3.5-38B	78.0	74.5	77.8	72.7	41.1	34.9	50.1	45.2
Qwen3-VL-8B	71.4	65.8	69.4	65.1	52.0	43.3	53.1	45.1
EuroLLM-22B	70.5	63.5	67.4	61.3	–	–	–	–
EuroLLM-9B	66.3	61.9	50.6	50.8	–	–	–	–
Qwen2.5-VL-7B	62.6	58.9	55.7	53.5	43.4	38.2	42.1	38.6
Qwen2.5-7B	61.7	57.4	61.7	55.4	–	–	–	–
Llama-3.1-8B	61.3	58.8	60.4	59.5	–	–	–	–
Mistral-7B	54.3	52.6	51.8	49.0	–	–	–	–
Gemma-3-4B	54.3	50.7	57.7	52.2	19.1	16.0	21.3	19.8
Gemma-2-2B	46.4	41.7	43.4	41.8	–	–	–	–
OpenAI models								
GPT-5.4[†]	**93.7**	**92.4**	**94.2**	**93.0**	**86.0**	**79.0**	**90.1**	**85.8**
GPT-4.1	91.3	88.6	93.0	91.5	68.5	56.7	78.8	71.5
GPT-5.4 mini	89.3	87.3	91.1	90.3	69.9	59.3	73.5	62.1
GPT-4.1-mini	85.6	82.0	88.1	84.7	65.7	53.5	74.8	66.3
GPT-5.4 nano	74.7	67.7	81.7	75.7	39.6	33.2	57.7	45.0

Table 2: Task accuracy (%) on SWESPECQA, reported both as micro-average over questions and macro-average over the six exam sources. Text results use the 1,266 questions that do not require the original page; PDF results use all 1,500 questions. Direct and CoT denote prompting style. Best result per column is **bolded**. [†]GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated here.

format. Source-specific extraction details are provided in Appendix B and the accompanying software.

Answer keys also differ across sources. EU MCQ SE, Strama SE, and Tolkprov SE have separate answer keys that are parsed with PyMuPDF. The remaining sources require an LLM-assisted OCR pipeline. For UMU Läkare SE and UU Apotekare SE, the visible green answer checks are removed from the PDF inputs used for evaluation as described in Appendix C.

All parsers produce a common format containing the question text, optional context, answer alternatives, gold answer, source document, page number, question type, and whether the item depends on an original visual or layout element. We set this flag when the parsed textual representation is insufficient to answer the question correctly without viewing the source page. This includes figures, images, maps, diagrams, chemical structures, symbol sequences, charts, graphs, and tables whose original layout carries information. In particular, EU numerical items that require their source table or chart are treated as visual/layout-dependent even when individual values can be transcribed.

We manually checked all 1,500 questions against the original source material. During this verification, we identified and corrected extraction errors, mostly minor transcription errors involving medical and chemical terminology in the UMU Läkare and UU Apotekare exams. We also found several incomplete page mappings that omitted pages needed to answer the target question. All identified issues were corrected before the final experiments.

3.3 Question and Answer Types

SweSpecQA consists mainly of multiple-choice questions. EU MCQ SE, Strama SE, UMU Läkare SE, and UU Apotekare SE use single-answer MCQs, but the number of answer choices varies by source: Strama uses three labeled choices (`1/X/2`), while EU items range from four to six alternatives and the medical and pharmacy sources retain their source-specific option sets. Tolkprov SE contains both standard MCQs and cloze-style MCQs, each with four options. Förarintyg SE contains the most varied format, including single-answer MCQs, a small number of multi-answer questions, and true/false blocks; its MCQs typically use four to six options. Questions that require the original figure, chart, table layout, chemical structure, clinical image, nautical chart, or similar page element are excluded from the text track but kept in the PDF track.

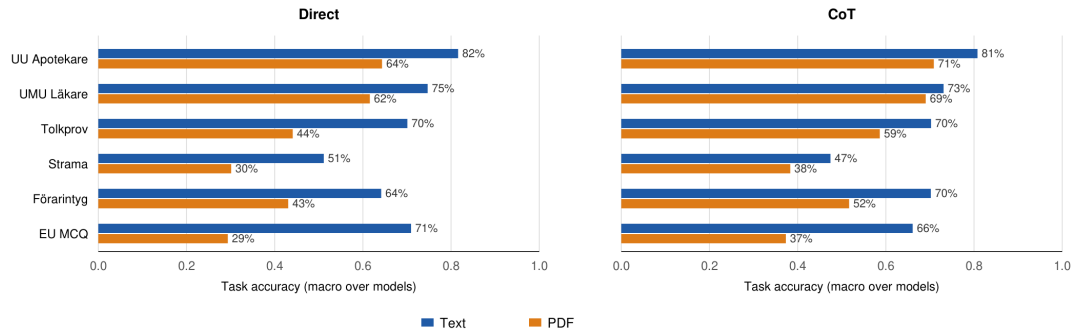


Figure 1: Task accuracy by exam, modality, and prompt type, macro-averaged over the models evaluated in each modality. Text bars use text-eligible items, PDF bars use all source items, and text-only models contribute only to the text bars. Matched text–PDF comparisons are reported separately in Section 5.3.

3.4 Evaluation Tracks

Inspired by multimodal exam benchmarks such as MMMU-Pro (Yue et al., 2025), we define two complementary evaluation tracks:

- **Text track:** Models receive the question text and answer alternatives extracted by our parsers. Only items answerable from this textual representation are included.
- **PDF track:** Models receive the full rendered source page or pages associated with the question. Departing from the vision track in MMMU-Pro, we do not provide a cropped image of the target question. **Instead, we provide the full mapped page or pages and the visible number of the target question.** The model must locate the relevant question in this page input, read it correctly, and then solve it. This track therefore evaluates document understanding more broadly, including layout interpretation, question localization, OCR, and problem solving. The layouts of the exam sources are illustrated in Appendix G.

Of the 1,500 questions, 234 require an original visual or layout-dependent source element and appear only in the PDF track; the remaining 1,266 are included in both tracks. The visual/layout-dependent subset comprises 79 EU MCQ, 25 Förarintyg, 26 UMU Läkare, and 104 UU Apotekare items; Strama and Tolkprov contain no such items in the benchmark.

4 Experimental Setup

We evaluate all models in a zero-shot, closed-book setting, without retrieval, browsing, or external tools. Each question is evaluated independently, without fine-tuning or few-shot examples on *SweSpecQA*. All instructions are written in Swedish.

4.1 Prompt Variants

Following the direct and reasoning-style evaluations used in the literature (Yue et al., 2024, 2025), we evaluate each track with two prompt variants.

- **Direct:** The model is instructed to return only the final answer in a fixed JSON format, without explanation.
- **Reasoning:** The model is instructed to reason step by step and place the final JSON answer on the last line.

Together with the two evaluation tracks, this gives four settings: `text-direct`, `text-cot`, `pdf-direct`, and `pdf-cot`. Each prompt is prefixed with a short exam-specific expert instruction, such as an instruction to answer as an expert in the respective domain. The exact expert prompts used in the experiments are provided in Appendix D.

4.2 Baseline Models

We evaluate 17 instruction-tuned models spanning OpenAI API models and open-weight models. The full list is provided in Table 6. Open-weight models are served locally with vLLM using deterministic decoding, while OpenAI models are evaluated through the official OpenAI API with temperature set to zero where supported. The PDF rendering settings, decoding hyperparameters, output limits, and scoring rules are provided in Appendix F. We include two EuroLLM models (Martins et al.,

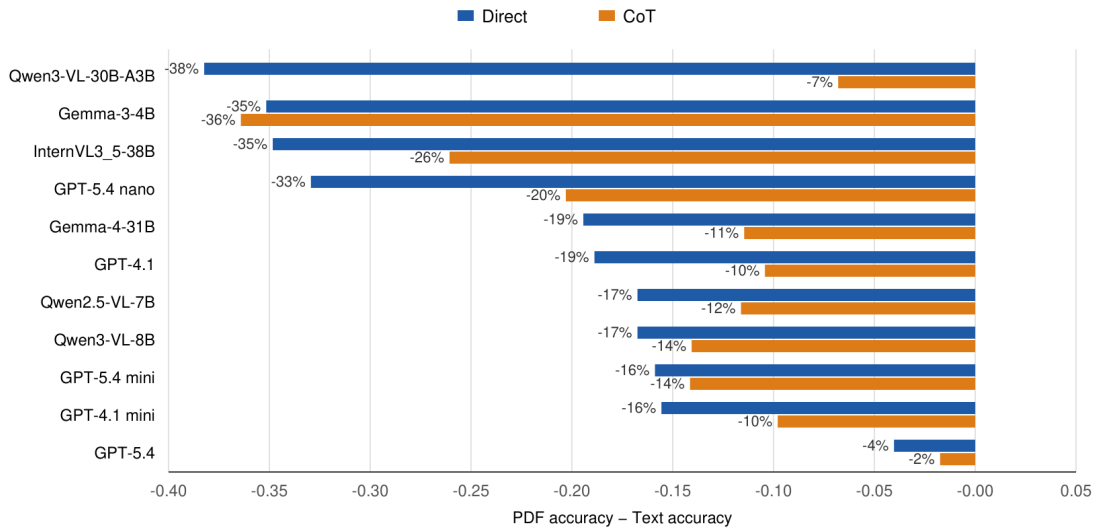


Figure 2: PDF-minus-text task-accuracy difference on the same 1,266 text-eligible questions. Negative values indicate lower accuracy when the question is presented on its full mapped source page(s).

2025; Ramos et al., 2026) because they are explicitly trained on all official EU languages including Swedish. The multimodal roster spans Qwen-VL, Gemma, InternVL, and GPT model families, including Qwen3-VL-8B, Qwen3-VL-30B-A3B, and InternVL3.5-38B. The PDF track is restricted to models with image input support; text-only models are evaluated on the text track only. We did not identify reliable human baselines or comparable official accuracy/pass-rate reference points for the exact public item sets included in *SweSpecQA*, so we do not report a human-performance number.

Models are asked to output their final answers as JSON objects at the end of their responses. Outputs that lack a recoverable answer are counted as incorrect. We use task accuracy as the main measure of answer correctness. If a model returns the full text of an answer choice rather than its label, we count it as correct only when it can be mapped unambiguously to the corresponding option. We also report *strict answer-label accuracy*, which requires the model to return the requested label directly. Since the six exams differ in size, we report both micro-average accuracy over questions and macro-average accuracy over exams.

5 Results

Table 2 reports both task micro-accuracy over questions and macro-average accuracy over the six exam sources for all models across the applicable evaluation settings. We discuss *strict answer-label accuracy* separately in Section 5.5. The detailed breakdown of each model's performance across exams is provided in Appendix H. Text-track results use the text-eligible subset; PDF-track results use all 1,500 items. Paired text–PDF comparisons use the same text-eligible subset in Figure 2. We additionally report paired bootstrap confidence intervals for the text–PDF gap and the effect of CoT prompting in Appendices I and J.

5.1 Overall Performance

GPT-5.4 is strongest under both micro and macro averaging in all four settings: its micro accuracies are 93.7%/94.2% on text and 86.0%/90.1% on PDF, while its macro accuracies are 92.4%/93.0% on text and 79.0%/85.8% on PDF. Macro averaging gives each exam source equal weight and changes some model rankings. On text-direct, Gemma-4-31B reaches 89.5% macro accuracy, ahead of GPT-4.1 at 88.6%, reversing their micro ordering (89.7% versus 91.3%). On text-CoT, GPT-5.4 mini reaches 90.3% macro accuracy versus 89.9% for Gemma-4-31B, again reversing their very close micro ordering (91.1% versus 91.2%). Thus, GPT-5.4 remains the strongest model and PDF input remains consistently harder, although the ordering of some models changes under macro averaging.

Among open-weight models, Gemma-4-31B is strongest overall on text under both micro and macro averaging, with 89.7%/91.2% micro and 89.5%/89.9% macro accuracy under direct/CoT prompting. Its PDF scores are 66.7%/76.8% micro and 51.4%/68.3% macro. The VLM results show

that the PDF difficulty is not limited to smaller models: InternVL3.5-38B reaches 78.0% micro accuracy on text-direct but 41.1% on PDF-direct, while Qwen3-VL-30B-A3B reaches 79.0% but 39.6%. Qwen3-VL-8B is comparatively more robust in direct PDF mode at 52.0%. Among text-only open-weight models, EuroLLM-22B remains strongest on text-direct by micro accuracy at 70.5% (63.5% macro).

5.2 Performance by Domain

Figure 1 shows descriptive accuracy by exam and input modality, macro-averaged over the models evaluated in each modality, while Table 8 in Appendix H reports per-model values. Text bars use each source's text-eligible items, whereas PDF bars use all source items; text-only models also contribute only to the text bars. The figure is therefore descriptive rather than a direct text–PDF comparison on the same questions; the matched question-level text–PDF comparison is reported separately in Section 5.3. Domain difficulty remains substantial. For GPT-5.4, UMU Läkare and UU Apotekare are the strongest domains, with text accuracies around 95–99% and PDF accuracies above 92% under direct prompting. EU MCQ shows a large aggregate text–PDF difference for GPT-5.4: 91.2%/85.3% on text versus 56.6%/70.8% on PDF. Strama is also comparatively difficult, especially for weaker VLMs. The pattern is not explained by answer format alone: most items are multiple choice, but performance varies markedly by source, question subtype, and document structure.

The temporal and subtype breakdown in Appendix L.2 further shows that within-source variation can be large. In EU MCQ, abstract logical reasoning remains harder than reading comprehension; UMU clinical cases differ from standard MCQs; and UU parts differ in difficulty. These effects motivate reporting both aggregate and source-level results.

5.3 Text–PDF Performance Gap

To isolate the cost of document processing, we compare text and PDF performance on the exact same 1,266 questions that do not require an original visual or layout element. Figure 2 reports the paired item-level PDF-minus-text difference. The result is systematic: every evaluated multimodal model performs worse from the full PDF page in both prompting conditions, and every paired 95% confidence interval is below zero. Under direct prompting, the drop ranges from -4.0 points for GPT-5.4 to -38.2 for Qwen3-VL-30B-A3B; InternVL3.5-38B and Gemma-3-4B also lose 34.8 and 35.2 points. Under CoT, the penalty ranges from -1.7 points for GPT-5.4 to -36.4 for Gemma-3-4B. Even the strongest model therefore shows a measurable PDF performance drop, although it is much smaller than for the open-weight VLMs.

Because the same 1,266 questions are used in both conditions, the gap cannot be explained by differences in the mix of visual questions. Instead, it reflects the combined difficulty of locating the target question, reading the page, identifying the answer choices and their labels, and solving the question from the original document rather than from cleaned text.

5.4 Visual and Layout-Dependent Questions

The PDF track contains 234 questions for which the original page carries information needed to answer correctly. This includes nautical charts, diagrams, plots, chemical structures, clinical images, symbol sequences, and layout-dependent tables. These items are excluded from the text track and evaluated only in the PDF track.

Difficulty varies strongly by source. GPT-5.4, for example, reaches 39%/65% on the 79 EU visual/layout-dependent items, 72%/68% on the 25 Förarintyg items, 77%/81% on the 26 UMU Läkare items, and 83%/88% on the 104 UU Apotekare items under direct/CoT prompting. The large spread indicates that "visual" is not a single capability: abstract pattern completion, table/chart reasoning, navigation symbols, clinical images, and chemical structures impose different demands.

The open-weight VLMs also vary substantially across sources and prompting styles, rather than showing the same pattern of difficulty on visual questions. Overall, the subset confirms that the PDF track measures capabilities absent from clean-text evaluation: locating the relevant information on the page, interpreting its spatial or visual structure, and combining it with the Swedish question text. The full model-by-source breakdown is shown in Figure 4 in Appendix H.

5.5 Reasoning and Answer Format

The effect of CoT varies substantially across models and between text and PDF input. In the PDF track, several models improve strongly with CoT: the largest gain is +20.6 points for Qwen3-VL-30B-A3B, while GPT-5.4 nano gains +18.1 points

	Option shuffle		Prefix	
Model	Direct	CoT	Pre	Post
GPT-4.1	90.3 ± 1.1	91.7 ± 1.4	12.8	15.0
GPT-5.4	92.2 ± 0.9	93.2 ± 1.2	10.9	9.2
GPT-5.4 nano	71.3 ± 1.9	77.5 ± 1.0	11.1	11.1
Llama-3.1-8B	64.7 ± 2.4	66.8 ± 2.3	11.8	10.7
Gemma-4-31B	87.6 ± 0.2	89.7 ± 0.2	8.5	9.2
EuroLLM-22B	69.3 ± 0.4	68.8 ± 1.0	11.4	10.7

Table 3: Robustness controls for six evaluated models. Option-shuffle values are task accuracy (%) with SD over three shuffled runs. Prefix values are ROUGE-L F1 (%) on the balanced pre- and post-cutoff samples.

Error type	Gemma	Qwen	GPT-mini	Total
Reasoning	7	14	2	23
Page/visual	7	3	2	12
Answer/output	2	2	2	6
Localization	3	1	1	5
Unclear	1	0	13	14
Total	20	20	20	60

Table 4: Manual error analysis of 60 incorrect PDF-CoT predictions, with 20 examples each from Gemma-4-31B, Qwen3-VL-30B-A3B, and GPT-5.4 mini.

and several other models gain around 9–10 points. By contrast, the effect is small for Qwen3-VL-8B (+1.1) and slightly negative for Qwen2.5-VL-7B (-1.3). On clean text, CoT can have the opposite effect, reducing accuracy by as much as 15.6 points. Thus, CoT is not a general accuracy boost; its effect depends strongly on both the model and whether the input is text or PDF.

A separate issue is whether models return the requested answer label. The gap between task accuracy and strict answer-label accuracy can be large: GPT-4.1 reaches 68.5%/78.8% task accuracy on PDF direct/CoT but only 15.5%/23.3% strict label accuracy. Other models, including GPT-5.4 mini and Gemma-4-31B, show much smaller gaps. This indicates that answer correctness and following the requested answer format can differ substantially.

We report whole-response JSON validity separately in Table 11. This is a formatting diagnostic rather than a measure of answer correctness, since reasoning prompts allow explanatory text before the final JSON answer.

5.6 Robustness and Contamination Controls

Because the source exams are publicly available, training-data contamination cannot be ruled out. We therefore use three complementary diagnostics. First, an option-shuffle control permutes the answer alternatives while moving the gold label with the correct option. Across the six models, shuffled-option accuracy remains broadly stable, with standard deviations of at most 2.4 percentage points. Second, we perform a prefix-completion test for verbatim memorization. We provide the first half of each sampled question text and ask the model to continue it. None of the six models reproduces the complete reference continuation exactly. Mean ROUGE-L F1 ranges from 8.5% to 15.0% across the model and pre-/post-cutoff combinations. Table 3 summarizes these experiments.

As an additional check, we examine PDF-direct accuracy by publication year and exam part. We do not observe a common sharp performance change across models and sources. These checks provide no clear evidence of simple memorization effects, but they cannot establish that contamination is absent, particularly because the source materials are public and cutoff information is incomplete for several models. Full diagnostic details and the temporal breakdown are reported in Appendix L.

5.7 PDF Error Analysis

To gain more insight into the failure modes in document-native evaluation, we manually analyzed a small sample of 60 incorrect PDF-CoT question–response pairs from three multimodal baselines: Gemma-4-31B, Qwen3-VL-30B-A3B, and GPT-5.4 mini, with 20 examples per model. For each example, we inspected the model's saved reasoning and final answer together with the corresponding exam page(s) to understand the source of the failure. We assigned a failure type only when the available evidence was sufficient to identify the underlying error. Fourteen cases for which there was not enough evidence to reliably determine the source of the failure were classified as *unclear*.

Among the 46 classifiable errors, half were reasoning or domain-knowledge errors (23/46), where the model identified the correct question and appeared to read the relevant information correctly, but made an error in its reasoning or lacked the domain knowledge needed to reach the correct answer. The remaining errors were related to document processing: 12 involved incorrect reading or interpretation of page content, six were answer-association or output errors, and five were question-localization errors. These failures included, for example, misreading values or relations in a table or figure, deriving the correct answer but returning

the wrong visible option label, and answering a neighboring question on the same page instead of the target question.

6 Conclusion

We introduced *SweSpecQA*, a Swedish-origin benchmark for specialized question answering across six institutional and professional domains. The benchmark pairs an extracted text track with a full-page PDF track, allowing us to separate clean-text question answering from document processing. Across 17 models, the strongest systems perform very well on extracted text, especially on medical and pharmaceutical exams, but document-native evaluation remains harder. Most importantly, on the matched text-eligible subset, every evaluated multimodal model suffers a consistent PDF performance drop. The PDF track additionally exposes visual and layout challenges that vary by source and substantial differences in how reliably models follow the requested answer format. CoT can recover part of the PDF performance gap for several models, including open-weight VLMs, but its effect varies strongly across models and between text and PDF input and can hurt text performance. These results show that clean-text evaluation can substantially overestimate reliability on original documents and that preprocessing remains an important component of Swedish exam-style QA systems.

Limitations

SweSpecQA is limited to Swedish examination and certification materials that are publicly available and covers only six domains. The dataset is not balanced across domains: physician and pharmacist exams make up a large share of the 1,500 questions, while other sources are smaller. Because the source exams are public, data contamination cannot be fully ruled out. Some models may have seen parts of the material during training or post-training; the temporal analysis is only a coarse diagnostic and cannot prove that the benchmark is contamination-free.

The visual/layout-dependent subset contains 234 questions (15.6% of the benchmark) and is unevenly distributed: EU MCQ, Förarintyg, UMU Läkare, and UU Apotekare contribute such items, while Strama and Tolkprov do not. It also combines several different kinds of visual content, including charts, tables, diagrams, chemical structures, clinical images, nautical symbols, and abstract visual patterns. Aggregate visual results should therefore be interpreted together with the source-level breakdown.

Strong benchmark scores should not be interpreted as evidence of real-world clinical or pharmaceutical competence. Passing closed-book exam questions does not imply safe deployment or the ability to consult and apply current official guidelines in practice. Additionally, we used GPT-5.4 for OCR-like transcription of complex PDFs and also evaluate it as a baseline model. Construction and evaluation were performed in separate calls, and evaluation prompts did not include information from the construction stage.

References

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. 2025. From KMMLU-redux to pro: A professional Korean benchmark suite for LLM evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19067–19096, Suzhou, China. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuanqi Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Al-

mubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Murathan Kurfalı, Shorouq Zahra, Evangelia Gogoulou, Luise Dürlich, Fredrik Carlsson, and Joakim Nivre. 2025. SweSAT-1.0: The Swedish University entrance exam as a benchmark for large language models. In *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, Tallinn, Estonia. University of Tartu Library.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and et al. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Birger Moëll, Fabian Farestam, and Jonas Beskow. 2025. Swedish medical llm benchmark: development and evaluation of a framework for assessing large language models in the swedish medical domain. *Frontiers in Artificial Intelligence*, 8:1557920.

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, Tórshavn, Faroe Islands. University of Tartu Library.

Miguel Moura Ramos, Duarte M Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M Guerreiro, Ricardo Rei, Nicolas Boizard, and et al. 2026. Eurollm-22b: Technical report. *arXiv preprint arXiv:2602.05879*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shengbang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

A Detailed Source Descriptions

This appendix records the source-specific details, selection rules, and exclusions summarized in Section 3.1.

EU MCQ SE. Practice materials for the EPSO (European Personnel Selection Office) aptitude test, distributed by UHR (Universitets- och högskolerådet, the Swedish Council for Higher Education) to candidates preparing for civil-service positions in EU institutions.[3] The exam covers three reasoning subdomains: verbal reasoning (*läsförståelse*), numerical reasoning (*tolkning av sifferuppgifter*), and abstract logical thinking (*logiskt tänkande*). All items are single-answer multiple-choice questions. The verbal reasoning items require drawing inferences from short passages; the numerical items require interpreting tables and graphs under time pressure; the logical items test non-verbal pattern completion. The benchmark includes UHR practice sets from 2016, 2017, 2021, and 2024.

[3] https://www.uhr.se/internationella-mojligheter/arbete-och-praktik-inom-eus-institutioner/infor-EU-uttagningsprov/Ovningsmaterial/

Förarintyg SE. Practice examinations for the Swedish inshore boating certificate (*Förarintyg*), published by NFB (Nämnden för båtlivsutbildning, the Swedish Council for the Education and Training of Yachtsmen).[4] The certificate is governed by Transportstyrelsen and tests theoretical knowledge required to navigate a recreational vessel in Swedish coastal waters within three nautical miles of a sheltered location. The exam covers COLREGs and Swedish maritime rules of the road (*sjövägsregler*), maritime safety (*sjösäkerhet*), and nautical chart reading. We retain only questions with fixed answers, including single-answer MCQ, multiple-correct-answer MCQ, and true/false blocks. Open-answer items are excluded from the benchmark. Several questions require interpreting navigation charts, lights, or symbols and therefore depend on visual information. NFB publishes separate exam sets for the East and West coastal regions with largely overlapping content; we include only the East-coast set to avoid near-duplicate items.

Strama SE. Annual educational quizzes published by Strama Västra Götaland for Swedish healthcare professionals.[5] The quizzes are distributed to Swedish healthcare professionals as continuing education and assess knowledge of national antibiotic treatment guidelines (*nationella behandlingsrekommendationer*), indication-based prescribing, antimicrobial resistance mechanisms, and infection epidemiology. All items are single-answer MCQs, often preceded by a short patient vignette or clinical scenario.

Tolkprov SE. Practice examinations for the authorized interpreter credential (*auktoriserad tolk*), administered by Kammarkollegiet (the Swedish Legal, Financial and Administrative Services Agency).[6] The written proficiency exam consists of three components; we include the first two. The realia component (Del 1) contains 60 single-answer MCQ items with four alternatives each, testing knowledge of Swedish civic institutions across five topic areas: social policy and society (10 items), labor market and education (10), healthcare and welfare (15), law and migration (20), and interpreter ethics and technique (5). The Swedish language proficiency component (Del 2) contains 60 four-option language questions: two 20-blank cloze passages and 20 dialogue-completion items. This tests lexical and grammatical competence in Swedish as the working language of interpretation. The third component (terminology translation from Swedish into the candidate's language) is excluded because it is language-pair-specific and lacks a fixed answer key for Swedish. Each component has a time limit of 60 minutes.

UMU Läkare SE. Theoretical proficiency examinations for physicians trained outside the EU/EEA, administered by Umeå University under commission from Socialstyrelsen (the Swedish National Board of Health and Welfare).[7] Passing the theoretical component is a prerequisite for the subsequent practical examination and, ultimately, for obtaining a Swedish medical licence. The theoretical exam is divided into an MCQ component, a clinical cases component, and a scientific article component; we include the first two and exclude the article component, as the target articles are in English. Items span internal medicine, surgery, pediatrics, obstetrics and gynecology, psychiatry, and pharmacology, and are grounded in Swedish clinical practice guidelines and the Swedish healthcare system. Several questions include radiological images, ECGs, microscopy slides, or dermatological photographs and are marked as requiring visual information.

UU Apotekare SE. Theoretical proficiency examinations for pharmacists trained outside the EU/EEA, administered by Uppsala University under commission from Socialstyrelsen.[8] As with the physician exam, passing the theoretical part is a requirement before the practical examination and Swedish licensure. The exam is structured into Part A and Part B, covering pharmaceutical chemistry, pharmacokinetics, pharmacodynamics, drug interactions, dispensing, and strict adherence to Swedish pharmaceutical regulations (*författningar*). We include both parts and exclude the scientific article component, which tests the understanding of an academic article in English. Questions that reference chemical structures, molecular

[4] https://batlivsutbildning.se

[5] https://strama.se

[6] https://www.kammarkollegiet.se/vara-tjanster/tolk/tolkproven

[7] https://www.umu.se/utbildning/anmalan-och-antagning/du-som-har-utlandsk-utbildning/kunskapsprov-for-lakare/

[8] https://www.uu.se/en/study/higher-education-in-sweden/foreign-educational-background/working-as-a-pharmacist-in-sweden/proficiency-test

diagrams, plots and similar are flagged as visual questions.

B Parser Details

This section gives a source-specific overview of the parsers used to construct *SweSpecQA*. All parsers produce the same output format: question text, optional context, answer alternatives, gold answer, source document, page number, question type, and a visual/layout-dependence flag. The resulting benchmark composition is summarized in Table 1 in the main paper, and representative PDF layouts are shown in Figure 3.

EU MCQ SE. The EU MCQ parser processes the Swedish question PDFs together with their corresponding answer keys. Questions are extracted from the PDF pages and converted into the common multiple-choice format. Question extraction and answer extraction are kept separate so that the answer key is not used when transcribing the questions. The parser also identifies questions that depend on figures, charts, diagrams, images, or table layout and marks them as visual-required. The official answer label is preserved even when the alternatives cannot be represented faithfully as plain text.

Förarintyg SE. The Förarintyg parser processes the question and answer-key PDFs separately. Since the exams contain several question formats, including single-choice, multi-select, true/false, chart questions, and navigation symbols, each page is parsed to recover the question type, alternatives, part and question number, and whether visual information is required. Answers are then matched to questions using their part and question numbers. Questions that require a chart, nautical symbol, route, map, or other visual element are marked as visual-required.

Strama SE. The Strama parser is mostly deterministic. It uses PDF or HTML text extraction, removes source-specific headers and layout artifacts, and segments questions using the characteristic 1/X/2 answer format. Patient cases and context shared across questions are preserved and attached to the relevant question. Gold answers are extracted from detailed answer keys when available, or from short-answer key documents using rules based on the answer-key layout. Since Strama questions are generally text-based, items are marked visual-required only when they explicitly refer to a figure or image.

Tolkprov SE. The Tolkprov parser uses deterministic layout extraction for the Kammarkollegiet practice material. For the realia section, it groups PDF text lines by visual row and reconstructs question stems and alternatives. For the language-proficiency section, it reconstructs the two long cloze passages and separately parses the 20 dialogue-completion items, producing one four-option item for each blank or dialogue. Gold answers are extracted from the answer tables by detecting the red-highlighted answer cells around each answer label.

UMU Läkare SE and UU Apotekare SE. The UMU Läkare parser renders each target page as an image, with the preceding page included as context when needed to cover cases where a question spans two pages. The parser extracts clinical cases, question stems, options, visual flags, and green-check answer labels into the common format. It returns a question only when the answerable portion is present on the target page. Questions requiring visuals such as ECGs, skin images, diagrams, or other non-text components are marked as visual-required.

The UU Apotekare parser processes the Part A and Part B PDFs page by page. Each page is rendered and parsed to extract question stems, options, visual flags, and green-check answer labels. As with UMU Läkare, questions that depend on chemical structures, molecular figures, plots, or other visual components are marked as visual-required.

For both exams, the gold answer labels are read from the source checks, while the visible green checks are removed from the PDF evaluation inputs as described in Appendix C.

C Removal of Visible Answer Markings

The UMU Läkare and UU Apotekare exams are originally published as single PDFs in which the correct options are marked with green check marks. These marks are useful for answer extraction, but they would leak the correct answer in the PDF track. We therefore create separate redacted copies of these exams exclusively for model evaluation.

We remove the check marks with a PyMuPDF-based redaction pipeline. The script reads the graphics layer of each page, finds small green vector drawings, and filters them by color, size, area, and page position so that it targets the answer-tick

pattern. The redaction is then applied only to the vector graphics layer, while text and embedded images are left unchanged. Across the UMU and UU items, the pipeline removes 1,010 expected green answer marks; a rendered-pixel check confirms zero residual green pixels in the redacted regions.

D Prompting Details

Table 5 lists the exam-specific expert prompts used in all evaluation settings. The same expert prompt is used for both the text and PDF tracks. The direct and CoT prompt variants are described in Section 4.1.

Key	Swedish Original	English Translation
eu_mcq_se	Du är expert på svenska EU-kunskapsfrågor och praktisk samhällskunskap.	You are an expert on Swedish EU knowledge questions and practical civics.
forarintyg_se	Du är expert på Förarintyg, navigation, sjövägsregler och svenskt båtliv.	You are an expert in inshore boating certificate, navigation, maritime regulations and Swedish boating.
strama_se	Du är expert på svenska Strama-riktlinjer, antibiotikabehandling och infektionsmedicin.	You are an expert on Swedish Strama guidelines, antibiotic treatment, and infectious medicine.
tolkprov_se	Du är expert på svensk auktorisationstolkning, tolketik och samhällsterminologi.	You are an expert on Swedish authorized interpretation, interpreter ethics, and civic terminology.
umu_lakare_se	Du är expert på svenska läkarlegitimationsfrågor, klinisk medicin och evidensbaserad vård.	You are an expert on Swedish medical licensing issues, clinical medicine, and evidence-based care.
uu_apotekare_se	Du är expert på svenska apotekarfrågor, farmakologi, farmaci och läkemedelsvetenskap.	You are an expert on Swedish pharmacist issues, pharmacology, pharmacy, and pharmaceutical science.

Table 5: Expert prompts used in the experiments with English translations.

E Model Details

Table 6 gives the full model identifiers used in the experiments. Text-only models are evaluated only on the text track, while multimodal models are evaluated on both the text and PDF tracks.

Model identifier	Provider	Modality	Data cutoff
Open-weight models			
Llama-3.1-**8B**-Instruct	Meta	Text	Dec. 2023
Mistral-**7B**-Instruct-v0.3	Mistral AI	Text	n/d
Qwen2.5-**7B**-Instruct	Alibaba	Text	n/d
Qwen2.5-VL-**7B**-Instruct	Alibaba	Multimodal	n/d
Qwen3-VL-**8B**-Instruct	Alibaba	Multimodal	n/d
Qwen3-VL-**30B**-A3B-Instruct	Alibaba	Multimodal	n/d
InternVL3.5-**38B**	OpenGVLab	Multimodal	n/d
Gemma-2-**2B**-it	Google	Text	n/d
Gemma-3-**4B**-it	Google	Multimodal	Aug. 1, 2024
Gemma-4-**31B**-it	Google	Multimodal	Jan. 2025
EuroLLM-**9B**-Instruct-2512	Utter Project	Text	1 December 2025
EuroLLM-**22B**-Instruct-2512	Utter Project	Text	1 December 2025
OpenAI models			
GPT-4.1-mini-2025-04-14	OpenAI	Multimodal	Jun. 1, 2024
GPT-4.1-2025-04-14	OpenAI	Multimodal	Jun. 1, 2024
GPT-5.4-nano-2026-03-17	OpenAI	Multimodal	Aug. 31, 2025
GPT-5.4-mini-2026-03-17	OpenAI	Multimodal	Aug. 31, 2025
GPT-5.4-2026-03-05[†]	OpenAI	Multimodal	Aug. 31, 2025

Table 6: Full model identifiers of the 17 baseline models evaluated on *SweSpecQA*. [†]GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated as a baseline. *n/d* indicates that no documented training-data cutoff was available.

F Evaluation Implementation Details

This appendix gives the implementation details needed to reproduce the reported evaluation settings. All evaluations are zero-shot: each question is evaluated independently, without examples from *SweSpecQA* in the prompt.

PDF rendering. For the PDF track, each model call contains exactly one target question and one or more full-page images according to that item's page mapping. We do not crop the target question and do not provide the extracted question text or answer options. The prompt specifies only the visible question number for the page input. Pages are rendered with PyMuPDF using a scale factor of 2.2 and saved as PNG images. The PDF track contains 779 rendered pages. For exams whose public PDFs contain visible answer markings, we use the redacted PDFs described in Appendix C.

Prompt format. Each prompt begins with the exam-specific expert instruction in Appendix D. In the text track, the prompt contains the parsed question and answer alternatives. In the PDF track, the prompt contains only the visible question number.

All prompts end with an answer-format instruction requiring the key `ratt_svar`. The required output format is:

```
{"ratt_svar": "..."}
```

For single-answer multiple-choice and cloze questions, `ratt_svar` should be a single option label. For multiple-answer questions, it should be a JSON list of labels. For true/false blocks, it should be a JSON object mapping each statement label to `S` or `F`. In the PDF track, models are further instructed to return the visible option label rather than the option text; if labels are not visible, they are instructed to assume the visual order `A`, `B`, `C`, `D`, `E`.

Setting	Value
Evaluation	Zero-shot
PDF image scale factor	2.2
PDF image format	PNG
PDF page input	Full rendered page(s)
Question crop	None
Questions per model call	One question per call
`text_direct` max output tokens	50
`text_reasoning` max output tokens	8192
`pdf_direct` max output tokens	50
`pdf_reasoning` max output tokens	8192

Table 7: Evaluation rendering, decoding, and output-length settings used in the experiments.

Prompt variants and output limits. In the direct setting, models are instructed to return only the final JSON object without explanation. In the reasoning setting, models are instructed to reason step by step and place the final JSON object on the last line, with no text after it. The maximum output length is 50 tokens for direct variants and 8192 tokens for reasoning variants.

Total Budget. All local model experiments were run on a single NVIDIA H100 GPU and completed in under 10 hours in total. Benchmark construction and OpenAI API experiments cost approximately USD 25.

Answer parsing and scoring. The parser allows several common output variants so that superficial formatting differences are not counted as task errors. It accepts valid or fenced JSON, the last valid JSON object in a longer response, and short answers such as a bare option label.

Outputs that do not contain a recoverable answer are counted as incorrect. For single-answer multiple-choice questions, the parsed answer must match the gold label. For multiple-answer questions, the predicted set of labels must exactly match the gold set; for true/false blocks, all required sub-answers must match. If a model returns the full option text instead of the option label, we map it back to a label only when the match is unambiguous. Task accuracy accepts these unambiguous mappings. **Strict answer-label accuracy** is task accuracy restricted to answers that did not require mapping the answer text back to its label. It therefore measures whether the model directly returns the requested label, rather than JSON formatting. Whole-response JSON validity is reported in Table 11.

G PDF Layout Examples

Figure 3 illustrates the layout variation in the PDF track. Models are given the full mapped rendered page or pages rather than a cropped question image or the extracted text. The examples cover the 11 question subtypes described in Section 3.1.

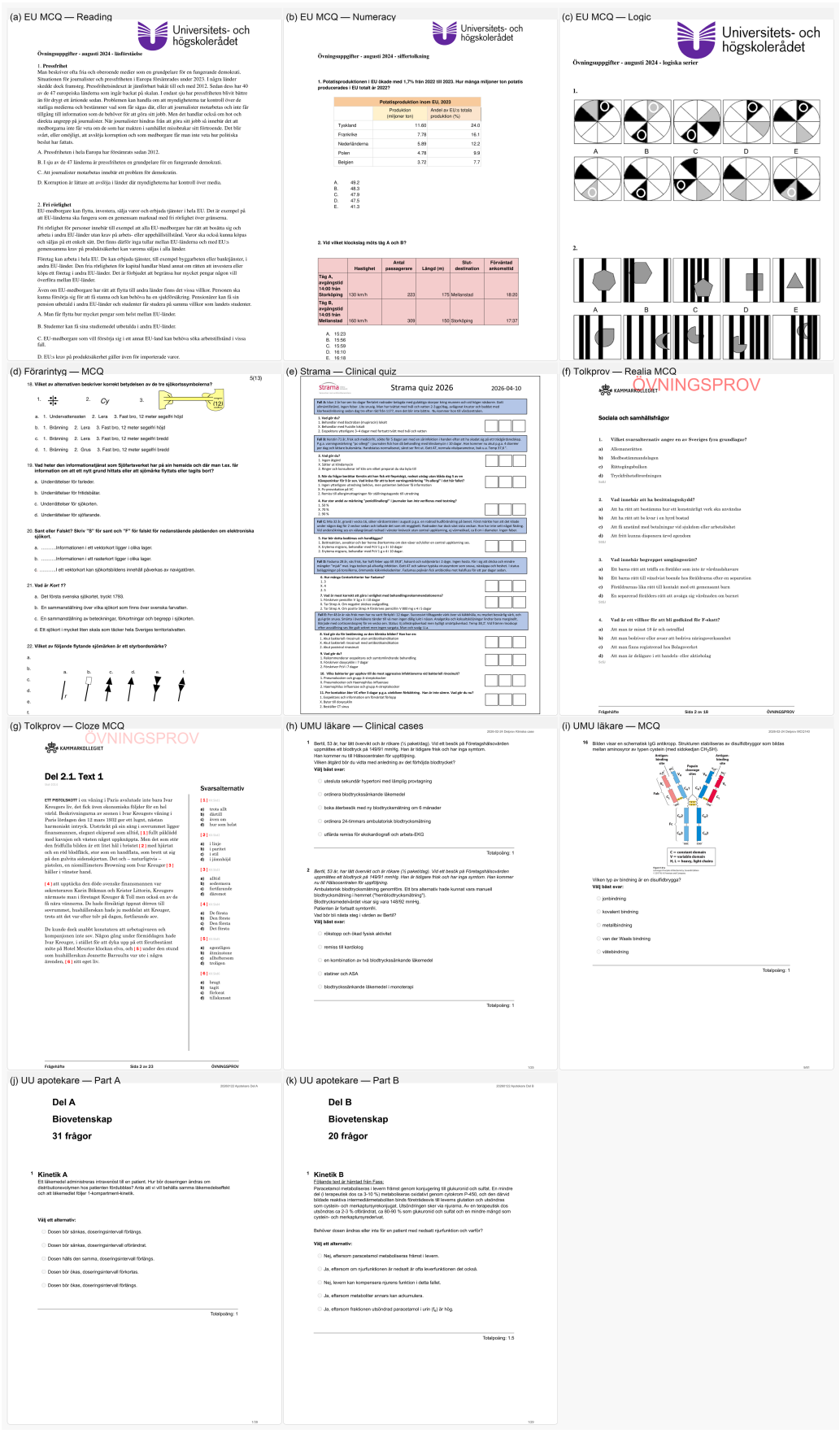


Figure 3: Representative PDF pages for the 11 question subtypes in *SweSpecQA*: EU reading, numeracy, and logic; Förarintyg multiple choice; Strama clinical quiz; Tolkprov realia and language-proficiency multiple choice; UMU clinical cases and standard multiple choice; and UU Apotekare Part A and Part B.

H Full Per-Exam Results

Table 8 gives the full per-exam breakdown behind Table 2 and the domain-level discussion in Section 5.2. The table reports task accuracy by model, exam, input track, and prompt variant.

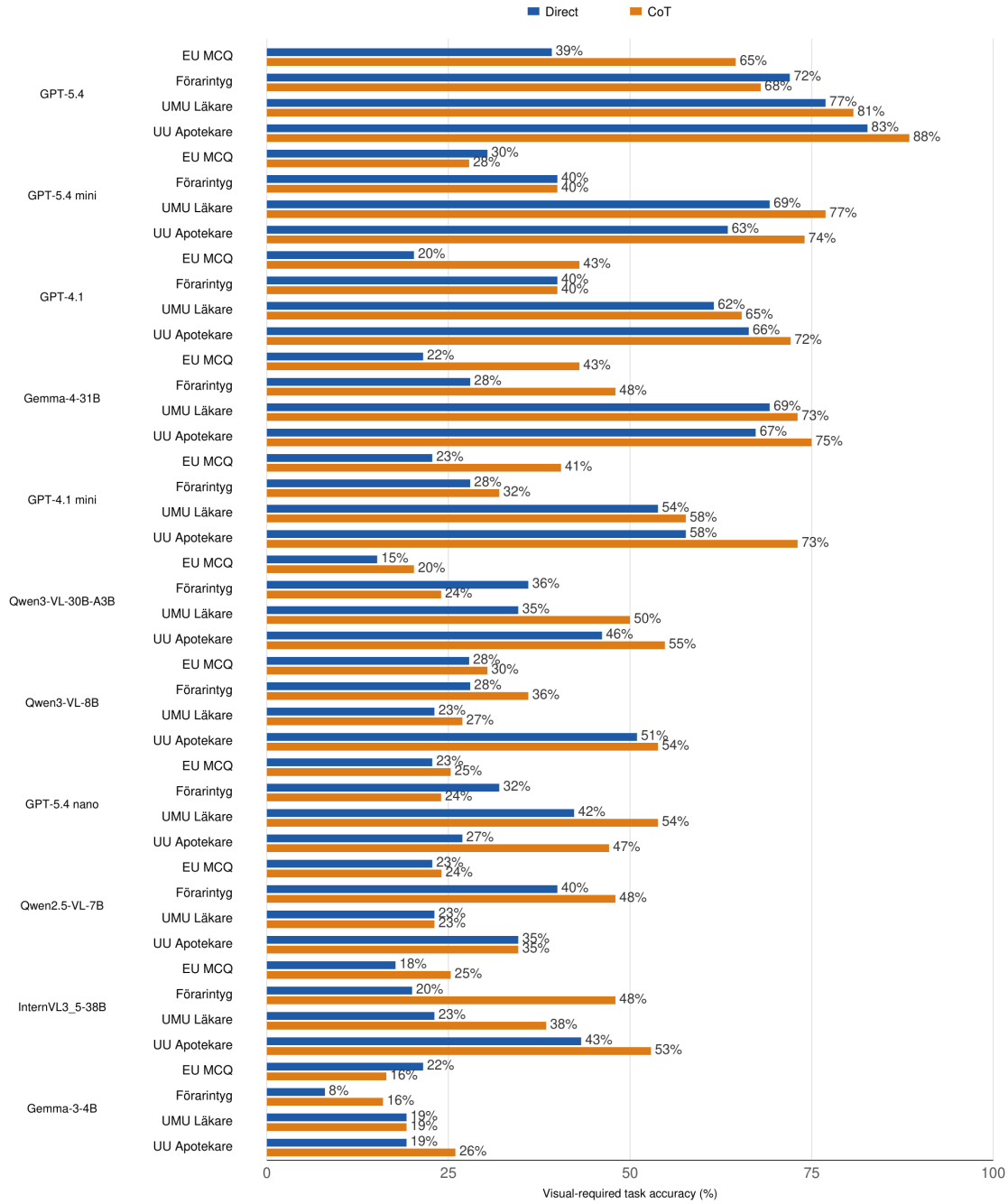


Figure 4: Task accuracy on the visual/layout-dependent subset in the PDF track, broken down by source and prompting style.

		Text		PDF				Text		PDF	
Model	Exam	D	CoT	D	CoT	Model	Exam	D	CoT	D	CoT
GPT-5.4†	EU MCQ	91.2	85.3	56.6	70.8	EuroLLM-22B	EU MCQ	52.9	58.8	–	–
	Förarintyg	93.8	100.0	77.2	78.9		Förarintyg	56.2	53.1	–	–
	Strama	79.0	81.5	75.5	78.0		Strama	43.0	35.0	–	–
	Tolkprov	96.7	97.5	79.2	98.3		Tolkprov	72.5	70.0	–	–
	UMU Läkare	95.2	96.0	92.1	93.5		UMU Läkare	73.1	72.7	–	–
	UU Apotekare	98.5	97.8	93.3	95.2		UU Apotekare	83.3	78.3	–	–
	Macro	92.4	93.0	79.0	85.8		*Macro*	63.5	61.3	–	–
GPT-4.1	EU MCQ	85.3	82.4	25.7	46.0	EuroLLM-9B	EU MCQ	64.7	52.9	–	–
	Förarintyg	87.5	100.0	56.1	64.9		Förarintyg	53.1	53.1	–	–
	Strama	77.0	78.5	29.5	54.0		Strama	42.5	37.0	–	–
	Tolkprov	91.7	95.8	64.2	90.8		Tolkprov	65.8	55.0	–	–
	UMU Läkare	93.3	94.6	83.2	87.1		UMU Läkare	69.7	49.1	–	–
	UU Apotekare	96.8	98.0	81.4	86.3		UU Apotekare	75.3	57.6	–	–
	Macro	88.6	91.5	56.7	71.5		*Macro*	61.9	50.8	–	–
Gemma-4-31B	EU MCQ	94.1	91.2	29.2	48.7	Qwen2.5-VL-7B	EU MCQ	70.6	52.9	28.3	30.1
	Förarintyg	93.8	90.6	42.1	73.7		Förarintyg	46.9	59.4	38.6	43.9
	Strama	68.0	70.5	9.0	19.0		Strama	48.5	42.0	28.5	28.0
	Tolkprov	93.3	96.7	58.3	88.3		Tolkprov	50.8	45.0	35.0	35.8
	UMU Läkare	92.3	94.4	82.6	89.5		UMU Läkare	61.4	54.3	47.7	45.3
	UU Apotekare	95.5	96.0	86.9	90.9		UU Apotekare	75.3	67.3	50.9	48.3
	Macro	89.5	89.9	51.4	68.3		*Macro*	58.9	53.5	38.2	38.6
GPT-5.4 mini	EU MCQ	85.3	88.2	41.6	38.1	Qwen2.5-7B	EU MCQ	64.7	50.0	–	–
	Förarintyg	90.6	96.9	45.6	47.4		Förarintyg	46.9	53.1	–	–
	Strama	70.5	71.0	48.0	56.5		Strama	43.0	42.0	–	–
	Tolkprov	90.0	95.8	60.8	63.3		Tolkprov	54.2	49.2	–	–
	UMU Läkare	92.7	93.9	81.0	81.8		UMU Läkare	61.8	63.3	–	–
	UU Apotekare	94.8	96.0	78.6	85.3		UU Apotekare	74.1	75.1	–	–
	Macro	87.3	90.3	59.3	62.1		*Macro*	57.4	55.4	–	–
GPT-4.1-mini	EU MCQ	82.4	79.4	26.5	39.8	Llama-3.1-8B	EU MCQ	70.6	70.6	–	–
	Förarintyg	81.2	87.5	47.4	57.9		Förarintyg	46.9	56.2	–	–
	Strama	58.5	62.0	34.5	49.5		Strama	44.5	32.0	–	–
	Tolkprov	85.0	90.8	55.8	83.3		Tolkprov	60.0	65.8	–	–
	UMU Läkare	90.2	92.3	78.8	81.6		UMU Läkare	65.3	61.8	–	–
	UU Apotekare	94.5	96.0	78.2	85.7		UU Apotekare	65.6	70.8	–	–
	Macro	82.0	84.7	53.5	66.3		*Macro*	58.8	59.5	–	–
Qwen3-VL-30B-A3B	EU MCQ	82.4	67.6	14.2	25.7	Mistral-7B	EU MCQ	52.9	55.9	–	–
	Förarintyg	71.9	81.2	40.4	42.1		Förarintyg	50.0	40.6	–	–
	Strama	48.0	34.0	37.0	35.5		Strama	39.5	28.0	–	–
	Tolkprov	79.2	72.5	25.0	54.2		Tolkprov	57.5	54.2	–	–
	UMU Läkare	82.7	72.0	38.4	68.5		UMU Läkare	53.2	51.8	–	–
	UU Apotekare	90.3	86.8	50.9	72.9		UU Apotekare	62.6	63.6	–	–
	Macro	75.7	69.0	34.3	49.8		*Macro*	52.6	49.0	–	–
InternVL3.5-38B	EU MCQ	85.3	73.5	23.9	28.3	Gemma-3-4B	EU MCQ	44.1	38.2	18.6	19.5
	Förarintyg	75.0	75.0	33.3	59.6		Förarintyg	56.2	53.1	10.5	17.5
	Strama	44.0	38.0	23.5	27.5		Strama	31.0	41.0	0.5	7.0
	Tolkprov	68.3	73.3	32.5	40.0		Tolkprov	51.7	54.2	20.0	27.5
	UMU Läkare	82.9	83.9	46.7	58.6		UMU Läkare	57.4	58.2	20.4	21.0
	UU Apotekare	91.8	92.3	49.3	56.8		UU Apotekare	63.6	68.6	25.9	26.5
	Macro	74.5	72.7	34.9	45.2		*Macro*	50.7	52.2	16.0	19.8
Qwen3-VL-8B	EU MCQ	73.5	70.6	32.7	34.5	Gemma-2-2B	EU MCQ	41.2	47.1	–	–
	Förarintyg	56.2	68.8	42.1	47.4		Förarintyg	31.2	40.6	–	–
	Strama	47.0	32.5	32.0	33.5		Strama	33.0	29.5	–	–
	Tolkprov	59.2	60.0	30.8	30.8		Tolkprov	42.5	38.3	–	–
	UMU Läkare	74.1	74.7	55.6	55.8		UMU Läkare	45.1	41.8	–	–
	UU Apotekare	85.0	84.3	66.7	68.3		UU Apotekare	57.4	53.6	–	–
	Macro	65.8	65.1	43.3	45.1		*Macro*	41.7	41.8	–	–
GPT-5.4 nano	EU MCQ	64.7	58.8	25.7	29.2						
	Förarintyg	53.1	84.4	40.4	35.1						
	Strama	52.5	52.0	13.5	33.0						
	Tolkprov	73.3	80.0	23.3	32.5						
	UMU Läkare	79.5	87.9	50.9	76.8						
	UU Apotekare	83.0	91.3	45.5	63.4						
	Macro	67.7	75.7	33.2	45.0						

Table 8: Task accuracy (%) by model, exam, and evaluation setting. *D* denotes direct prompting and *CoT* reasoning prompting; dashes indicate that a text-only model was not evaluated on PDF input. Macro is the unweighted average over the six exams. †GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated here.

I Text–PDF Gap Analysis

This section supports the text–PDF comparison in Section 5.3. Table 9 reports paired bootstrap confidence intervals for the PDF-minus-text accuracy difference on the matched text-eligible subset, matched by question ID within each model and prompt type. We resample matched question pairs with replacement, recompute the accuracy difference, and report the 2.5th and 97.5th percentiles.

Model	Prompt	Pairs	Text	PDF	PDF–Text [95% CI]	PDF wins	Text wins	Ties
GPT-4.1	Direct	1266	91.3	72.4	-18.9 [-21.2, -16.5]	22	261	983
GPT-4.1-mini	Direct	1266	85.6	70.1	-15.6 [-18.0, -13.1]	44	241	981
GPT-5.4†	Direct	1266	93.7	89.7	-4.0 [-5.7, -2.4]	29	80	1157
GPT-5.4 mini	Direct	1266	89.3	73.5	-15.9 [-18.5, -13.3]	48	249	969
GPT-5.4 nano	Direct	1266	74.7	41.8	-32.9 [-36.0, -29.9]	55	472	739
InternVL3.5-38B	Direct	1266	78.0	43.2	-34.8 [-37.9, -31.7]	60	501	705
Qwen2.5-VL-7B	Direct	1266	62.6	45.9	-16.7 [-19.8, -13.7]	110	322	834
Qwen3-VL-30B-A3B	Direct	1266	79.0	40.8	-38.2 [-41.3, -35.2]	54	538	674
Qwen3-VL-8B	Direct	1266	71.4	54.7	-16.7 [-19.5, -14.0]	77	289	900
Gemma-3-4B	Direct	1266	54.3	19.1	-35.2 [-38.5, -31.8]	86	531	649
Gemma-4-31B	Direct	1266	89.7	70.2	-19.4 [-22.0, -16.9]	32	278	956
GPT-4.1	CoT	1266	93.0	82.6	-10.4 [-12.5, -8.4]	28	160	1078
GPT-4.1-mini	CoT	1266	88.1	78.3	-9.8 [-12.0, -7.6]	43	167	1056
GPT-5.4†	CoT	1266	94.2	92.5	-1.7 [-3.1, -0.4]	26	48	1192
GPT-5.4 mini	CoT	1266	91.1	76.9	-14.1 [-16.5, -11.8]	43	222	1001
GPT-5.4 nano	CoT	1266	81.7	61.4	-20.3 [-23.1, -17.5]	63	320	883
InternVL3.5-38B	CoT	1266	77.8	51.7	-26.1 [-29.1, -22.9]	78	408	780
Qwen2.5-VL-7B	CoT	1266	55.7	44.1	-11.6 [-14.9, -8.1]	176	323	767
Qwen3-VL-30B-A3B	CoT	1266	70.9	64.1	-6.8 [-9.7, -3.9]	130	216	920
Qwen3-VL-8B	CoT	1266	69.4	55.4	-14.1 [-17.1, -11.1]	105	283	878
Gemma-3-4B	CoT	1266	57.7	21.3	-36.4 [-39.7, -33.2]	67	528	671
Gemma-4-31B	CoT	1266	91.2	79.7	-11.5 [-13.6, -9.4]	32	177	1057

Table 9: Paired bootstrap confidence intervals for the item-level micro PDF-minus-text accuracy delta on the 1,266 questions that do not require the original page.Pairs are matched by question ID within each model and prompt type. †GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated here.

J Reasoning Prompt Analysis

This section supports the CoT analysis in Section 5.5. Table 10 reports paired bootstrap confidence intervals for the CoT-minus-direct accuracy difference, matched by question ID within each model and input modality.

Model	Input	Pairs	Direct	CoT	CoT–Direct [95% CI]	CoT wins	Direct wins	Ties
GPT-4.1	PDF	1500	68.5	78.8	+10.3 [+8.3, +12.3]	206	52	1242
GPT-4.1-mini	PDF	1500	65.7	74.8	+9.1 [+6.8, +11.3]	225	89	1186
GPT-5.4†	PDF	1500	86.0	90.1	+4.1 [+2.5, +5.8]	111	49	1340
GPT-5.4 mini	PDF	1500	69.9	73.5	+3.7 [+1.6, +5.7]	153	98	1249
GPT-5.4 nano	PDF	1500	39.6	57.7	+18.1 [+15.6, +20.7]	357	85	1058
InternVL3.5-38B	PDF	1500	41.1	50.1	+9.0 [+6.4, +11.7]	275	140	1085
Qwen2.5-VL-7B	PDF	1500	43.4	42.1	-1.3 [-2.7, +0.1]	47	67	1386
Qwen3-VL-30B-A3B	PDF	1500	39.6	60.2	+20.6 [+17.6, +23.7]	455	146	899
Qwen3-VL-8B	PDF	1500	52.0	53.1	+1.1 [-0.5, +2.9]	95	78	1327
Gemma-3-4B	PDF	1500	19.1	21.3	+2.2 [-0.2, +4.6]	194	161	1145
Gemma-4-31B	PDF	1500	66.7	76.8	+10.1 [+8.1, +12.1]	198	47	1255
GPT-4.1	Text	1266	91.3	93.0	+1.7 [+0.3, +3.2]	53	31	1182
GPT-4.1-mini	Text	1266	85.6	88.1	+2.4 [+0.7, +4.3]	79	48	1139
GPT-5.4†	Text	1266	93.7	94.2	+0.6 [-0.6, +1.7]	29	22	1215
GPT-5.4 mini	Text	1266	89.3	91.1	+1.7 [+0.2, +3.3]	64	42	1160
GPT-5.4 nano	Text	1266	74.7	81.7	+7.0 [+4.6, +9.4]	164	76	1026
InternVL3.5-38B	Text	1266	78.0	77.8	-0.2 [-2.2, +1.7]	81	84	1101
Qwen2.5-7B	Text	1266	61.7	61.7	+0.0 [-2.6, +2.6]	143	143	980
Qwen2.5-VL-7B	Text	1266	62.6	55.7	-7.0 [-9.5, -4.4]	95	183	988
Qwen3-VL-30B-A3B	Text	1266	79.0	70.9	-8.1 [-10.4, -5.8]	62	165	1039
Qwen3-VL-8B	Text	1266	71.4	69.4	-2.0 [-4.1, +0.2]	87	112	1067
Gemma-2-2B	Text	1266	46.4	43.4	-3.0 [-5.5, -0.6]	108	146	1012
Gemma-3-4B	Text	1266	54.3	57.7	+3.5 [+0.9, +6.0]	153	109	1004
Gemma-4-31B	Text	1266	89.7	91.2	+1.5 [+0.1, +2.9]	51	32	1183
Llama-3.1-8B	Text	1266	61.3	60.4	-0.9 [-3.6, +1.9]	147	158	961
Mistral-7B	Text	1266	54.3	51.8	-2.5 [-5.1, +0.1]	119	151	996
EuroLLM-22B	Text	1266	70.5	67.4	-3.2 [-5.6, -0.8]	99	139	1028
EuroLLM-9B	Text	1266	66.3	50.6	-15.6 [-18.7, -12.5]	118	316	832

Table 10: Paired bootstrap confidence intervals for the item-level micro CoT-minus-direct task-accuracy delta. Pairs are matched by question ID within each model and input modality. †GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated here.

K Answer-Format Diagnostics

Table 11 compares task accuracy with strict answer-label accuracy. Task accuracy uses the forgiving parser described in Appendix F; strict label accuracy excludes answers that are correct only after mapping the answer text back to its label. JSON validity is reported as a separate diagnostic.

Model	Track	N	Task	Strict label	Gap	Option text	JSON
GPT-5.4†	Text-D	1266	93.7	93.7	0.0	0.0	100.0
GPT-5.4†	Text-CoT	1266	94.2	94.2	0.0	0.0	0.0
GPT-5.4†	PDF-D	1500	86.0	61.1	24.9	26.7	100.0
GPT-5.4†	PDF-CoT	1500	90.1	64.8	25.3	25.9	0.0
GPT-4.1	Text-D	1266	91.3	91.3	0.0	0.0	40.3
GPT-4.1	Text-CoT	1266	93.0	93.0	0.0	0.0	0.0
GPT-4.1	PDF-D	1500	68.5	15.5	53.0	60.1	99.4
GPT-4.1	PDF-CoT	1500	78.8	23.3	55.5	59.9	0.3
GPT-5.4 mini	Text-D	1266	89.3	89.3	0.0	0.0	100.0
GPT-5.4 mini	Text-CoT	1266	91.1	91.1	0.0	0.0	0.0
GPT-5.4 mini	PDF-D	1500	69.9	67.8	2.1	2.7	99.9
GPT-5.4 mini	PDF-CoT	1500	73.5	72.0	1.5	2.0	47.5
GPT-4.1-mini	Text-D	1266	85.6	85.6	0.0	0.0	0.0
GPT-4.1-mini	Text-CoT	1266	88.1	88.1	0.0	0.0	0.0
GPT-4.1-mini	PDF-D	1500	65.7	34.1	31.7	39.2	43.7
GPT-4.1-mini	PDF-CoT	1500	74.8	26.5	48.3	53.8	0.1
GPT-5.4 nano	Text-D	1266	74.7	74.7	0.0	0.0	100.0
GPT-5.4 nano	Text-CoT	1266	81.7	81.7	0.0	0.0	0.0
GPT-5.4 nano	PDF-D	1500	39.6	27.7	11.9	20.9	99.9
GPT-5.4 nano	PDF-CoT	1500	57.7	47.3	10.4	13.0	40.5
Gemma-4-31B	Text-D	1266	89.7	89.7	0.0	0.0	0.0
Gemma-4-31B	Text-CoT	1266	91.2	91.2	0.0	0.0	0.0
Gemma-4-31B	PDF-D	1500	66.7	66.2	0.5	0.9	0.0
Gemma-4-31B	PDF-CoT	1500	76.8	74.7	2.1	2.3	0.0
Qwen3-VL-30B-A3B	Text-D	1266	79.0	79.0	0.0	0.0	100.0
Qwen3-VL-30B-A3B	Text-CoT	1266	70.9	70.9	0.0	0.0	0.0
Qwen3-VL-30B-A3B	PDF-D	1500	39.6	21.4	18.2	41.7	99.9
Qwen3-VL-30B-A3B	PDF-CoT	1500	60.2	36.1	24.1	30.2	0.1
InternVL3.5-38B	Text-D	1266	78.0	78.0	0.0	0.0	0.0
InternVL3.5-38B	Text-CoT	1266	77.8	77.8	0.0	0.0	0.0
InternVL3.5-38B	PDF-D	1500	41.1	10.4	30.7	43.8	0.0
InternVL3.5-38B	PDF-CoT	1500	50.1	13.9	36.2	44.7	0.0
Qwen3-VL-8B	Text-D	1266	71.4	71.4	0.0	0.0	100.0
Qwen3-VL-8B	Text-CoT	1266	69.4	69.4	0.0	0.0	0.0
Qwen3-VL-8B	PDF-D	1500	52.0	38.5	13.5	22.2	100.0
Qwen3-VL-8B	PDF-CoT	1500	53.1	35.9	17.2	27.1	81.5
Qwen2.5-VL-7B	Text-D	1266	62.6	62.6	0.0	0.0	99.1
Qwen2.5-VL-7B	Text-CoT	1266	55.7	55.7	0.0	0.0	0.0
Qwen2.5-VL-7B	PDF-D	1500	43.4	36.7	6.7	13.5	95.6
Qwen2.5-VL-7B	PDF-CoT	1500	42.1	36.7	5.3	10.6	67.4
Gemma-3-4B	Text-D	1266	54.3	54.3	0.0	0.0	0.0
Gemma-3-4B	Text-CoT	1266	57.7	57.7	0.0	0.0	0.0
Gemma-3-4B	PDF-D	1500	19.1	14.0	5.1	16.3	0.0
Gemma-3-4B	PDF-CoT	1500	21.3	17.4	3.9	13.5	0.3
Qwen2.5-7B	Text-D	1266	61.7	61.7	0.0	0.0	99.1
Qwen2.5-7B	Text-CoT	1266	61.7	61.7	0.0	0.0	1.6
EuroLLM-22B	Text-D	1266	70.5	70.5	0.0	0.0	99.9
EuroLLM-22B	Text-CoT	1266	67.4	67.4	0.0	0.0	0.0
EuroLLM-9B	Text-D	1266	66.3	66.3	0.0	0.0	54.3
EuroLLM-9B	Text-CoT	1266	50.6	50.6	0.0	0.0	0.0
Llama-3.1-8B	Text-D	1266	61.3	61.3	0.0	0.0	100.0
Llama-3.1-8B	Text-CoT	1266	60.4	60.4	0.0	0.0	0.0
Mistral-7B	Text-D	1266	54.3	54.3	0.0	0.0	99.9
Mistral-7B	Text-CoT	1266	51.8	51.8	0.0	0.0	0.0
Gemma-2-2B	Text-D	1266	46.4	46.4	0.0	0.1	0.0
Gemma-2-2B	Text-CoT	1266	43.4	43.4	0.0	0.0	0.0

Table 11: Item-level micro task accuracy versus strict answer-label accuracy by model and track. Task accuracy accepts an unambiguous option-text answer; strict label accuracy does not. The JSON column is a separate whole-response diagnostic and is not part of the strict-label definition. Gap is Task minus Strict label. †GPT-5.4 was used for LLM-assisted benchmark construction and is also evaluated here.

L Robustness and Contamination Diagnostics

L.1 Option-Shuffle and Prefix-Completion Controls

We run two complementary controls on the text track. The option-shuffle control samples 25 text-eligible questions from each of the six sources (150 total), permutes answer-option order with seeds 13, 29, and 47, and moves the gold label with the option text. Table 3 reports the mean and standard deviation across the three shuffled runs for direct and reasoning prompts.

The prefix-completion test samples 100 text-eligible questions per model, balanced as 50 before and 50 after the model's documented training-data cutoff. It gives the model the first half of the question and measures continuation overlap with the original second half.

L.2 Temporal Breakdown

Figure 5 provides the performance changes across years, as discussed in Section 5.6. We do not observe a common sharp increase or decrease in performance across years.

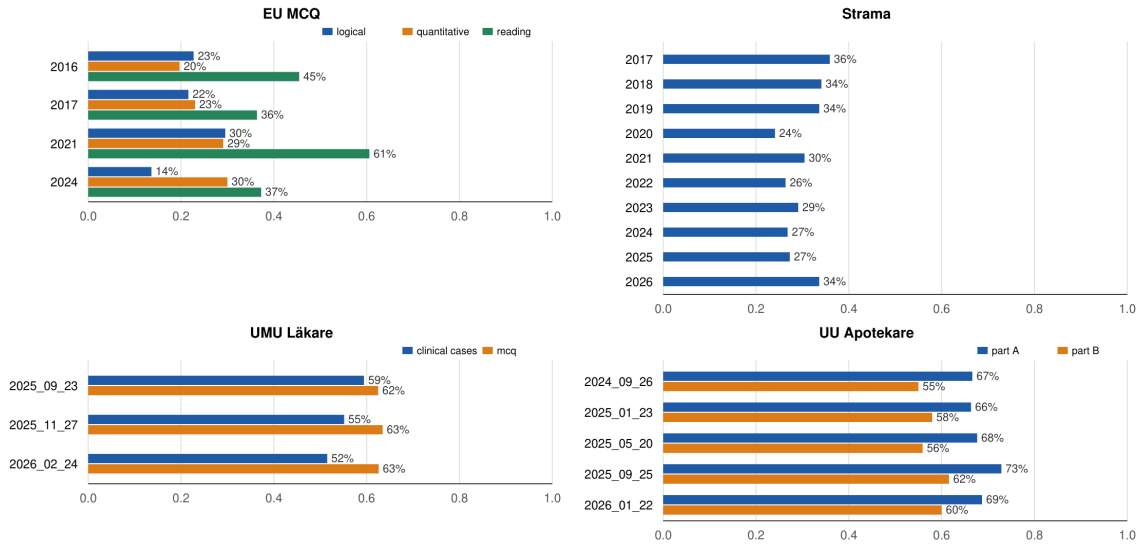


Figure 5: PDF-direct task accuracy by exam date/year and exam part, macro-averaged over models evaluated in the PDF track.