

ContextDril at ImageCLEF 2025 Multimodal Reasoning: Evaluating VLMs' Multimodal, Multilingual and Multidomain Reasoning Capabilities via Thinking Budget Variations and Textual Augmentation

Authors	Krazheva, Vasilena T.;Markova, Diana;Dimitrov, Dimitar I.;Koychev, Ivan;Nakov, Preslav
Citation	V. T. Krazheva, D. Markova, D. I. Dimitrov, I. Koychev, and P. Nakov, "ContextDril at ImageCLEF 2025 Multimodal Reasoning: Evaluating VLMs' Multimodal, Multilingual and Multidomain Reasoning Capabilities via Thinking Budget Variations and Textual Augmentation," CEUR Workshop Proc, 2025
Publisher	CEUR-WS
Download date	2026-06-15 04:45:34
Link to Item	https://hdl.handle.net/20.500.14634/1552

ContextDrift at ImageCLEF 2025 Multimodal Reasoning: Evaluating VLMs’ Multimodal, Multilingual and Multidomain Reasoning Capabilities via Thinking Budget Variations and Textual Augmentation*

Notebook for the ImageCLEF, Task 4 - MultimodalReason Lab at CLEF 2025

Vasilena T. Krazheva^{1,*}, Diana Markova^{1,*}, Dimitar I. Dimitrov¹, Ivan Koychev¹ and Preslav Nakov²

¹Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”, Bulgaria

²Mohamed Bin Zayed University of Artificial Intelligence, UAE

Abstract

With the growing capabilities of vision-language models (VLMs), current systems achieve impressive performance on tasks requiring the integration of vision and language, such as image captioning, simple visual question answering, and visual dialogue. However, it is often claimed that these models fall short when deeper reasoning is required. In this paper, we investigate this claim through the ImageCLEF 2025 MultimodalReasoning task, which challenges models to solve multiple-choice questions in image format across a number of subjects and languages. Using Gemini 2.0 Flash and 2.5 Flash, we study the effect of reasoning capacity and budget, external textual transcription, and prompt design on the EXAMS-V benchmark for Bulgarian and English. Our results indicate that, contrary to expectation, VLMs can perform remarkably well on multimodal reasoning tasks in both languages. In particular, they are able to solve tasks in Physics and Science with an accuracy of over 80%. We identify thinking budget as the main contributing factor. Additionally, we demonstrate a setting where unconstrained thinking budget might deteriorate performance in Biology and Chemistry. The system submitted ranked first on English and Bulgarian leaderboards with respective 89.65% and 90.50% accuracy scores.

Keywords

Multimodal Reasoning, Vision-Language Model, Gemini, Optical Character Recognition, Visual Question Answering

1. Introduction

Recent advances in VLMs have enabled new capabilities for solving problems that span both visual and textual modalities [1]. This multimodal reasoning ability is essential for a diverse set of applications such as document question answering, educational tutoring systems, and embodied intelligence. However, evaluating these systems remains challenging — especially when reasoning must occur across images, diagrams, and multiple languages.

Existing benchmarks for evaluating multimodal reasoning have provided valuable insights into the capabilities of VLMs for a range of tasks. One such benchmark is Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) [2], which consists of college-level exam questions, spanning a wide range of academic subjects and fields including mathematics, science, the humanities, and the arts. It has become the *de facto* benchmark for measuring the multimodal reasoning capabilities of VLMs.

The ImageCLEF 2025 MultimodalReasoning lab [3] at the Conference and Labs of the Evaluation Forum (CLEF) [4] addresses VLM evaluation in a broader and more language-inclusive manner by selecting the EXAMS-V dataset [5] for training and validation. It consists of 20,932 multiple-choice questions covering 20 school subjects across 11 languages and incorporates multimodal features such

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*These authors contributed equally.

✉ vasikr44@gmail.com (V. T. Krazheva); dianamarkovakn@gmail.com (D. Markova); ilijanovd@fmi.uni-sofia.bg (D. I. Dimitrov); koychev@fmi.uni-sofia.bg (I. Koychev); preslav.nakov@mbzuai.ac.ae (P. Nakov)

ORCID 0009-0003-5295-1609 (V. T. Krazheva)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

as text, images, tables, figures, diagrams, maps, scientific symbols, and equations. The competition also includes a held-out test set of 3,565 new questions in three additional languages (Urdu, Kazakh, and Spanish), introduced for the 2025 edition. Unlike other benchmarks, it provides a broader linguistic scope and places a particular emphasis on lower-resource languages [5]. The task itself is defined as follows: *given an image containing a multiple-choice question with three to five answer options and associated metadata, the objective is to identify the single correct answer.*

In this working notes paper, we investigate the performance of selected free-tier proprietary Gemini vision-language models on the ImageCLEF MultimodalReasoning task. We explore the impact of reasoning budget constraints, prompt design, and external Optical Character Recognition (OCR) textual transcription. Focusing on English and Bulgarian, we analyze performance trends across subjects and modalities. Our system ranked first on both English and Bulgarian leaderboards.

2. Related Work

Early approaches to Visual Question Answering (VQA) typically relied on heavily engineered modular architectures, where separate components handled image encoding, question interpretation, and answer classification. For tasks involving text within images, systems incorporated OCR pipelines to extract textual content, which was then fused with visual features for downstream reasoning [6, 7]. Recent advances in vision-language models have replaced such hand-crafted systems with unified transformer-based architectures that jointly model vision and language.

In fact, multimodal models have multiple encoders (for each modality) and then fuse the embeddings together to create a shared representation space; decoders operate over the shared latent space to produce output in the desired modality [1]. Examples of such models include GPT-4o [8], Claude 3.5 Sonnet [9], Gemini [10], Qwen2-VL-7B [11], LLaVA [12], and Gemma 3 [13].

Large Language Models (LLMs), such as OpenAI’s o1 [14], have dramatically improved performance on increasingly complex tasks by scaling test-time computation during problem-solving. The combination of multimodal capabilities and extended chain-of-thought fine-tuning and alignment has been recently implemented in models such as QVQ-72B-Preview [15], Kimi-VL-A3B-Thinking [16], and Gemini 2.5 [17], achieving SOTA results on the MMMU benchmark [18].

Prompt engineering has proven essential for extracting reasoning behavior from foundation models. As demonstrated in GPT-3 [19], few-shot prompting can enable models to generalize with minimal supervision in certain contexts. Furthermore, it has been shown that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks [20].

Our system builds on these advances in the following ways:

We select multimodal Gemini models as the core of our VQA system. Specifically, Gemini 2.5 Flash (thinking) serves as a primary inference component. Gemini 2.0 Flash (non-thinking) is employed in two roles: (1) as a baseline and experimental playground, and (2) to assess whether external textual transcription (via OCR) can enhance performance by better eliciting the model’s textual reasoning capabilities. Prompt engineering strategies are also considered to optimize performance.

3. Materials and Methods

3.1. Data

The dataset provided for the MultimodalReason task is an expanded version of EXAMS-V [5]. Each question is in image format and has corresponding metadata for language, subject, grade, presence of tables, figures, diagrams and chemical structures. Due to temporal and computational limitations, participation and further analysis are restricted to two language subsets. Namely, English and Bulgarian were selected as representatives of a higher-resource and a lower-resource language, respectively.

Table 1 shows the validation split count distribution of questions, grouped by language. *With Figure* refers to questions whose associated images contain a graphical element, while *Text Only* refers to

Table 1

Count Distribution of Text Only and With Figure questions per subject for English and Bulgarian questions in the validation sets.

Subject	English		Bulgarian	
	Text Only	With Figure	Text Only	With Figure
Biology	31	16	92	8
Chemistry	75	25	94	6
Physics	53	47	60	40
Science	97	3	–	–
Sociology	–	–	95	5
Total	256	91	341	59

questions whose image representations contain only text (*type* attribute).

3.2. Methodology

The main phases in our experimental workflow are data preparation, prompt engineering, model querying, and output post-processing & evaluation (Figure 1).

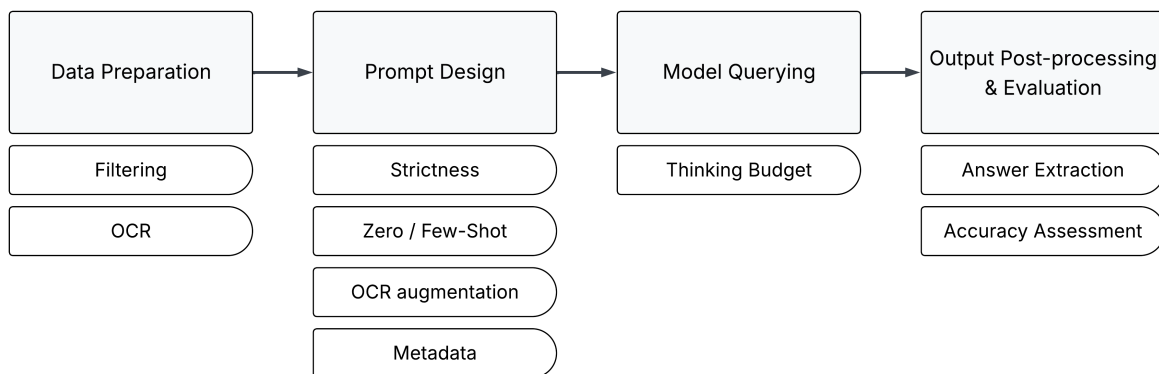


Figure 1: Experimental workflow.

3.2.1. Data Preparation

Preprocessing involved merging the dataset *parquet* files and filtering only the languages of interest. Additionally, for the Bulgarian validation set, answer keys were mapped to the corresponding unified English letters. Textual content extraction was performed with two OCR engines, each supporting both languages. The first, Tesseract OCR [21, 22], is an open-source OCR engine widely used in academic research. The second, OCRSpace [23], provides a cloud-based OCR service, often used in applied research.

3.2.2. Prompt Design

Prompt formatting can significantly affect the performance of evaluated models [24, 25]. Therefore, two fundamentally different prompting approaches were undertaken:

- **Approach 1:** A handcrafted task-specific prompt was designed as recommended in [26], utilizing the following techniques: role-play, step-by-step instructions (Chain of Thought) and contextualization.

- **Approach 2:** Meta prompting technique [27] was undertaken by instructing GPT-4o to generate and improve a prompt. The final version adheres to the Structured Prompt Template [28] by systematically organizing the prompt into the distinct components – task introduction, task detail, output format, few-shot examples and query. The prompt incorporates a structured input in *json* format.

Both prompt types integrate the question metadata provided in the dataset.

We hypothesize that augmenting the prompt with OCR text transcription could better engage the reasoning capabilities of the models and ensure focused problem understanding. To test this, prompts with and without external transcription were formed. Also, to measure the effectiveness of in-context model adaptation through samples, zero-shot and one-shot versions of the prompts were considered. Prompt versions and templates can be found in the Prompt 1 and Prompt 2 subsections of the Appendix section.

3.2.3. Model Querying

The specific release versions of Gemini 2.0 Flash and Gemini 2.5 Flash used are **gemini-2.0-flash** and **gemini-2.5-flash-preview-04-17**. All experiments were conducted using the following default configurations: temperature set to 1 and topP to 95. Gemini 2.5 Flash uses a default topK of 64, while 2.0 Flash defaults to a topK of 40.

Experiments were run with limited and unconstrained thinking budget. The thinking budget parameter¹ guides the model on the number of thinking tokens it can use when generating a response [29]. Higher values correspond to more detailed reasoning. By default, it is unconstrained; we denote this option with the ∞ symbol.

3.2.4. Output Post-processing & Evaluation

Task submission requires all answers to be one of the letters 'A', 'B', 'C', 'D' or 'E'. To ensure that the submission files adhered to the requirements, the following post-processing steps were applied to the models' responses: (1) extraction of the answer letter if it was not readily provided, and (2) mapping the answer symbol to one of the letters listed above when the model responded in the alphabet of the question's language. The official competition evaluation metric is *Accuracy*.

4. Experiments & Results

4.1. Pre-submission evaluations

A limited number of evaluations were performed with and without external OCR textual transcription, zero-shot and few-shot prompting, and different thinking budgets. Table 2 provides a summary of pre-submission runs. The best two runs for each language are in **bold**.

Since Gemini 2.0 Flash quota limitations were more favorable, and under the assumption that effects would be sufficiently similar when using Gemini 2.5 Flash, experiments with and without OCR augmentation were performed. OCR proved beneficial for Bulgarian with 2.0 Flash, contributing to a 10% increase in accuracy. Incorporating OCR data resulted in 95.25% accuracy for Bulgarian when using 2.5 Flash, and was thus used for submission runs.

However, for English, we observed a slight decrease in performance for Gemini 2.5 Flash runs with a 1024 thinking budget, and therefore chose to refrain from adding external textual transcription for the first submission run. These experiments were configured with a limited budget, originally motivated by shorter processing times.

Each last run per language in Table 2 was carried out with ∞ thinking budget and OCR augmentation, leading to longer execution times and substantial performance gains on the English set. Specifically, the

¹The Google GenAI API defines **thinkingBudget** as an integer in the range 0 to 24576

accuracy on the English validation set increased from 57.92% to 78.09%, potentially highlighting the impact ∞ budget has on model performance.

Table 2

Accuracy of pre-submission runs on validation sets.

Model	Prompt	Thinking Budget	OCR Type	Adaptations	Val
Bulgarian					
Gemini 2.0 Flash	Prompt 1	-	-	Zero-Shot	0.7875
Gemini 2.0 Flash	Prompt 1	-	Tesseract	Zero-Shot	0.8875
Gemini 2.5 Flash	Prompt 1	1024	Tesseract	Zero-Shot	0.9525
Gemini 2.5 Flash	Prompt 2	∞	OCRSpace	Few-Shot	<u>0.9675</u>
English					
Gemini 2.0 Flash	Prompt 1	-	-	Zero-Shot	0.4928
Gemini 2.5 Flash	Prompt 1	1024	Tesseract	Zero-Shot	0.5706
Gemini 2.5 Flash	Prompt 1	1024	-	Zero-Shot	0.5792
Gemini 2.5 Flash	Prompt 2	∞	OCRSpace	Few-Shot	<u>0.7809</u>

Due to nearing submission deadlines and quota restrictions, the best-performing systems were selected based on these preliminary validation results. Later experiments were conducted to explore a broader configuration space.

4.2. Official Submission Results

The official results of the competition are presented in Table 3. Our system achieved first place on both English and Bulgarian leaderboards with respective 89.65% and 90.50% accuracy scores.

Table 3

Leaderboard accuracy on test sets and ranking of submitted runs.

Model	Prompt	Thinking Budget	OCR Type	Adaptations	Test	Ranking
Bulgarian						
Gemini 2.5 Flash	Prompt 2	∞	OCRSpace	Few-Shot	0.9050	1st
Gemini 2.5 Flash	Prompt 1	1024	Tesseract	Zero-Shot	0.9050	1st*
English						
Gemini 2.5 Flash	Prompt 2	∞	OCRSpace	Few-Shot	0.8965	1st
Gemini 2.5 Flash	Prompt 1	1024	-	Zero-Shot	0.8086	4th*

**Since the authors participated with two different team names, which were subsequently united into one, there are two submissions per subtask.*

4.3. Post-submission investigation

To better understand the pronounced difference in accuracy for the English subtask and the unexpected consistency in accuracy for Bulgarian questions between the two approaches, we perform a series of ablations and modifications to the experimental settings. Tables 4 and 5 in the Appendix present validation accuracy for all experimental configurations.

4.3.1. External OCR contribution

OCR augmentation visibly improved accuracy in Gemini 2.0 Flash experiments for both languages. However, it is unclear whether the addition of external textual transcription contributes to the performance of 2.5 Flash on the validation sets.

OCR external transcription boosted accuracy in Gemini 2.0 Flash (non-thinking) experiments for Bulgarian (Table 4) with over 10% (experiments #15, #11, #7). For English (Table 5), we observe a smaller, but marked max performance increase of 7.2% (experiments #11, #15, #8).

Following further experiments for Bulgarian with Gemini 2.5 Flash, the highest accuracy achieved of 97.25% was in experiment #1 (No OCR) (Table 4). Experiment runs #2 (No OCR) and #13 (OCRSpace), with fixed otherwise settings, achieved corresponding scores of 97.00% and 96.75%, showing a decrease of 0.25% when OCR transcription is included. In contrast, when transcription is added to the setup of experiment run #3 (No OCR), we observe a max accuracy increase of 0.75% and a score of 97% (experiment run #12 with OCRSpace). All other groups of experiments show a max OCR boost of less than 0.75%. These fluctuations of 0.25-0.75% could be due to the generative nature of the model.

Trends on the English validation set are inconclusive (Table 5). On the one hand, experiments #3 (No OCR), #9 (Tesseract), #12 (OCRSpace) have accuracies of respectively 78.96%, 80.40%, 80.69%, indicating a max score increase of 1.73%. On the other hand, experiments #1 (No OCR) and #13 (OCRSpace), with accuracies of respectively 80.12% and 78.09%, show a decrease of 2.03%. This is also the case with experiments #6 (No OCR) and #10 (Tesseract) – we observe reduction (although smaller) of 0.86%. However, when using OCRSpace with the same settings (experiment #14), we note an increase of 1.45%.

Overall, for Gemini 2.5 Flash experiments, OCR external transcription did not result in consistent performance benefits. This could be attributed to Gemini 2.5 Flash’s superior visual understanding capabilities, reducing reliance on external textual transcriptions.

4.3.2. Zero-Shot vs. Few-Shot prompting

Few-shot prompting occasionally improved performance slightly, though as can be seen in Tables 4 and 5, results remain inconclusive.

Namely, for English (Table 5), experiments #1 (One-Shot) and #2 (Zero-Shot) with respective accuracies of 80.12% and 79.25% show a slight increase of 0.87%. However, experiments #3 (Zero-Shot) and #5 (One-Shot) with corresponding scores of 78.96% and 78.39%, demonstrate a small decrease of 0.57%.

For Bulgarian (Table 4), experiments #1 (One-Shot) and #3 (Zero-Shot) with scores 97.25% and 96.25% indicate an increase of 1%; experiments #2 (One-Shot) and #5 (Zero-Shot) follow the same trend with a marginal difference of 1%.

These minor differences may be due to the models already producing well-structured responses - in all cases, answer key extraction was practically reduced to a simple regular expression over the last five response characters.

4.3.3. Reasoning Budget Variations

In light of the stark difference in accuracy for experiments on the English validation set of around 20%, and the clearly noticeable, though smaller difference on the test set of less than 10%, we hereby analyze how thinking budget contributes to performance.

In order to isolate the effect of other improvements, which might be attributed to adaptation or external OCR text transcription, we conduct experiments with zero-shot prompting (Approach 1), no OCR augmentation, and thinking budgets of 1024, 8192 and ∞ (Table 5; experiments #6, #4 and #3, respectively). Corresponding scores achieved are 57.92%, 78.96% and 78.96%. Although the last two thinking budget settings yielded higher overall accuracy, Figure 2 reveals a more complex dynamic.

In the case of Physics and Science questions, the ∞ (unconstrained adaptive) configuration improved performance in both modalities. Specifically, for problems in Physics, overall accuracy increased from 60% (thinking budget 1024) to a score of 86% (thinking budget ∞). Similarly, overall Science accuracy increased to 96% (thinking budget ∞), while for 1024 thinking budget it was 53%. The higher limited

budget configuration of 8192 resulted in accuracy values between those achieved in 1024 and ∞ configurations. This is also true modality-wise (Figure 2).

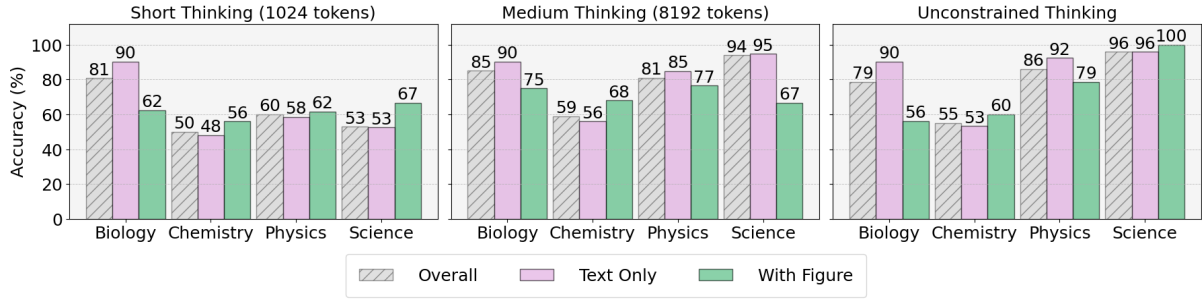


Figure 2: Accuracy per subject across three thinking budget scenarios.

However, the aforementioned dependency between thinking budget value and accuracy does not hold for Biology and Chemistry. In particular, for Biology questions, the score dropped from 81% with 1024 budget to 79% with ∞ budget, and reached its highest value of 85% when the parameter was set to 8192. Interestingly, the reduction in accuracy was for *With Figure* questions (Figure 2). We observe a similar trend for Chemistry, where overall accuracy peaked at 59% with 8192 thinking budget and reduced to 55% when thinking was unconstrained. Moreover, the reduction in accuracy for Chemistry was observed in both modalities (Figure 2).

This non-monotonicity is likely related to the underthinking and overthinking phenomena, suggesting the existence of optimal reasoning length. While this effect has been previously studied in LLMs [30, 31, 32, 33], the same pattern might naturally extend for multimodal reasoning tasks in VLMs.

5. Conclusion

In this working notes paper, we presented our results and analysis for the ImageCLEF 2025 MultimodalReasoning competition. By examining our pre-submission and post-submission experiments, we conclude that dataset questions vary in complexity — language-wise, subject-wise, and split-wise.

Regarding the proprietary models used, we found that reasoning capacity directly affects performance on both English and Bulgarian subsets. Remarkably, we discovered that Gemini 2.5 Flash performs better in the visual modality for certain subjects when the thinking budget is limited — potentially indicating a failure to self-calibrate its chain-of-thought reasoning length relative to problem demands. In addition, it is worth noting that external textual transcription substantially improved accuracy in 2.0 Flash experiments and occasionally resulted in slight increases in performance in 2.5 Flash experiments.

Nevertheless, we acknowledge that our analysis did not include experiments across all languages available in the competition dataset. As a result, it remains uncertain whether our findings generalize to other languages and problem formulations. Future work could be directed at investigating this extrapolation explicitly. Furthermore, response token-level metadata—such as prompt, thoughts, output, and total token counts—can be tracked to examine their relation to correctness. Gemini’s thought summaries option [34] could also be used to reveal the model’s internal problem-solving pathway for dataset problems.

Acknowledgments

The work is partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

Declaration on Generative AI

The authors have not employed any Generative AI tools in this work.

References

- [1] M. Noyan, S. Paniego, C. P. Gosthipaty, Aritra Roy, Vision Language Models (Better, faster, stronger), 2025. URL: <https://huggingface.co/blog/vlms-2025>.
- [2] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, W. Chen, MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI (2023). URL: <https://arxiv.org/pdf/2311.16502>. doi:10.1109/CVPR52733.2024.00913.
- [3] D. Dimitrov, M. S. Hee, Z. Xie, R. Joyti Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of ImageCLEF 2025 – Multimodal Reasoning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
- [4] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ștefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvey, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
- [5] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: <https://aclanthology.org/2024.acl-long.420>. doi:10.18653/v1/2024.acl-long.420.
- [6] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards VQA Models That Can Read, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June (2019) 8309–8318. URL: <https://arxiv.org/pdf/1904.08920>. doi:10.1109/CVPR.2019.00851.
- [7] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti, B. M. Team, ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data (2020). URL: <https://arxiv.org/pdf/2001.07966>.
- [8] OpenAI, GPT-4o System Card (2024). URL: <https://arxiv.org/pdf/2410.21276>.
- [9] Anthropic, Claude 3.5 Sonnet Model Card Addendum (2024). URL: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.
- [10] Google, Gemini 2.0 Flash - Model Card (2025). URL: <https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf>.
- [11] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution (2024). URL: <https://arxiv.org/pdf/2409.12191>.
- [12] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual Instruction Tuning, Advances in Neural Information Processing Systems 36 (2023). URL: <https://arxiv.org/pdf/2304.08485>.
- [13] Gemma Team, Gemma 3 Technical Report (2025). URL: <https://arxiv.org/pdf/2503.19786>.
- [14] OpenAI, OpenAI o1 System Card (2024). URL: <https://cdn.openai.com/o1-system-card-20241205.pdf>.

- [15] Qwen Team, QVQ: To See the World with Wisdom | Qwen, 2024. URL: <https://qwenlm.github.io/blog/qvq-72b-preview/>.
- [16] Kimi Team, Kimi-VL Technical Report (2025). URL: <https://arxiv.org/pdf/2504.07491>.
- [17] Google, Gemini 2.5 Flash Preview - Model Card (2025). URL: <https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash-preview.pdf>.
- [18] Y. Xiang, N. Yuansheng, Z. Kai, Z. Tianyu, MMMU Leaderboard, 2025. URL: <https://mmmu-benchmark.github.io/>.
- [19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, *Advances in Neural Information Processing Systems* 35 (2022). URL: <https://arxiv.org/pdf/2201.11903>.
- [21] R. Smith, An overview of the Tesseract OCR engine, *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR 2 (2007)* 629–633. URL: <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf>. doi:10.1109/ICDAR.2007.4376991.
- [22] GitHub, tesseract-ocr/tesseract: Tesseract open source ocr engine (main repository), 2025. URL: <https://github.com/tesseract-ocr/tesseract>.
- [23] 9t9 software GmbH, Free OCR API V2025, Online OCR, Searchable PDF Creator and OCR Software, 2025. URL: <https://ocr.space/>.
- [24] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate Before Use: Improving Few-Shot Performance of Language Models, *Proceedings of Machine Learning Research* 139 (2021) 12697–12706. URL: <https://arxiv.org/pdf/2102.09690>.
- [25] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, S. Hasan, Does Prompt Formatting Have Any Impact on LLM Performance? (2024). URL: <https://arxiv.org/pdf/2411.10541>.
- [26] H. He, M. Ye, J. Zhang, X. Cai, J. Liu, B. Du, D. Tao, Reasoning-OCR: Can Large Multimodal Models Solve Complex Logical Reasoning Problems from OCR Cues? (2025). URL: <https://arxiv.org/pdf/2505.12766>.
- [27] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncarenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, P. Resnik, The Prompt Report: A Systematic Survey of Prompt Engineering Techniques (2024). URL: <https://arxiv.org/pdf/2406.06608>.
- [28] Y. Liu, J. Xu, Li, L. Zhang, Q. Chen, X. Feng, Y. Chen, Z. Guo, Y. Yang, P. Cheng, Beyond Prompt Content: Enhancing LLM Performance via Content-Format Integrated Prompt Optimization (2025). URL: <https://arxiv.org/pdf/2502.04295>.
- [29] Google, Image understanding | Gemini API | Google AI for Developers, 2025. URL: <https://ai.google.dev/gemini-api/docs/image-understanding>.
- [30] X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, R. Wang, Z. Tu, H. Mi, D. Yu, Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs (2024). URL: <https://arxiv.org/pdf/2412.21187>.
- [31] Y. Wu, Y. Wang, M. Csail, T. Du, S. Jegelka, T. U. Munich, Y. Wang, When More is Less: Understanding Chain-of-Thought Length in LLMs (2025). URL: <https://arxiv.org/pdf/2502.07266>.
- [32] Y. Wang, Q. Liu, J. Xu, T. Liang, X. Chen, Z. He, L. Song, D. Yu, J. Li, Z. Zhang, R. Wang, Z. Tu, H. Mi, D. Yu, Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs (2025). URL: <https://arxiv.org/pdf/2501.18585>.

- [33] J. Su, J. Healey, P. Nakov, C. Cardie, *Between Underthinking and Overthinking: An Empirical Study of Reasoning Length and correctness in LLMs* (2025). URL: <https://arxiv.org/pdf/2505.00127>. doi:10.48550/arXiv.2505.00127.
- [34] Google, *Gemini API | Google AI for Developers*, 2025. URL: <https://ai.google.dev/gemini-api/docs>.

Appendix

Table 4

Accuracy on the Bulgarian validation set for all experiment runs.

No	OCR Type	Model	Prompt	Adaptations	Thinking Budget	Val Accuracy
1	No OCR	Gemini 2.5 Flash	Prompt 1	One-Shot	∞	0.9725
2	No OCR	Gemini 2.5 Flash	Prompt 2	One-Shot	∞	0.9700
12	OCRSpace	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.9700
8	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	8192 tokens	0.9700
13	OCRSpace	Gemini 2.5 Flash	Prompt 2	One-Shot	∞	0.9675 / 1st
9	Tesseract	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.9675
14	OCRSpace	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.9650
3	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.9625
4	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.9600
5	No OCR	Gemini 2.5 Flash	Prompt 2	Zero-Shot	∞	0.9600
10	Tesseract	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.9525 / 1st
15	OCRSpace	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.9025
6	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	0 tokens	0.8925
11	Tesseract	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.8875
7	No OCR	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.7875

Table 5

Accuracy on the English validation set for all experiment runs.

No	OCR Type	Model	Prompt	Adaptations	Thinking Budget	Val Accuracy
12	OCRSpace	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.8069
9	Tesseract	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.8040
1	No OCR	Gemini 2.5 Flash	Prompt 2	One-Shot	∞	0.8012
2	No OCR	Gemini 2.5 Flash	Prompt 2	Zero-Shot	∞	0.7925
3	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	∞	0.7896
4	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	8192 tokens	0.7896
5	No OCR	Gemini 2.5 Flash	Prompt 1	One-Shot	∞	0.7839
13	OCRSpace	Gemini 2.5 Flash	Prompt 2	One-Shot	∞	0.7809 / 1st
14	OCRSpace	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.5937
6	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.5792 / 4th
10	Tesseract	Gemini 2.5 Flash	Prompt 1	Zero-Shot	1024 tokens	0.5706
11	Tesseract	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.5648
15	OCRSpace	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.5187
7	No OCR	Gemini 2.5 Flash	Prompt 1	Zero-Shot	0 tokens	0.4986
8	No OCR	Gemini 2.0 Flash	Prompt 1	Zero-Shot	N/A	0.4928

Prompt 1

OCR data is conditionally included according to the experiment cases. We note that, for few-shot experiments with Prompt 1, the same example as in Prompt 2 is concatenated. Also, the content variable is substituted and formatted based on the question metadata.

Prompt 1 Template, English

```
PROMPT_TEMPLATE_STRICTER = (  
  "You are a sophisticated Vision-Language Model (VLM) capable of analyzing images  
  containing multiple-choice questions."  
  " To guide your analysis, you may adopt the following process:\n"  
  "0. Consider the subject of question is {subject} and image contains {content}.\n"  
  "1. Image Analysis: Examine the image closely, identifying key elements such as text,  
  diagrams, and any other relevant features.\n"  
  "-{ocr}\n"  
  "2. Question Text Extraction: Extract the text of the question\n"  
  "3. Extract Answer Choices: Identify and extract the answer choices provided in the image\n"  
  " - if the answer options are not enumerated with letters, do enumerate them with  
  letters (A, B, C, D, ...)\n"  
  "4. Look for additional visual elements such as tables, diagrams, charts, or graphs.\n"  
  "5. Ensure to consider any multilingual or multidomain aspects of the image, including text  
  in different languages or mathematical/physics/scientific notation.\n"  
  "6. Analyze the complete context and data provided\n"  
  "7. Select correct answer based solely on analysis.\n"  
  "8. Respond by only the corresponding letter (single capital letter) without any extra  
  explanation.\n"  
  "9. If the answer is not clear, still provide the best guess as single capital letter.\n\n"  
  "Always respond with a single capital letter (A, B, C, D, E) without any extra explanation."  
)
```

Prompt 1 Template, Bulgarian

```
PROMPT_TEMPLATE_BUL_STRICTER = (  
  "Ти си комплексен Vision-Language модел (VLM) способен да анализира изображения,  
  съдържащи multiple-choice questions."  
  " В насочването на анализите си, подходи така:\n"  
  "0. Взemi предвид, че предметът на въпроса е свързан с {subject} и изображението  
  съдържа {content}.\n"  
  "-{ocr}\n"  
  "1. Анализ на изображение: Изследвай отблизо изображението, идентифицирай  
  ключови елементи като текст, диаграми, и всякакви други релевантни характеристики.\n"  
  "2. Извечи текста, който представлява въпроса\n"  
  "3. Идентифицирай и извечи опциите за отговор на въпроса \n"  
  " - Ако отговорите не са номерирани с букви, номерирай ги с български букви  
  (А, Б, В, Г, Д)\n"  
  "4. Потърси допълнителни визуални елементи, като таблици, диаграми, графики  
  или фигури.\n"  
  "5. Увери се, че вземаш предвид всички многоезични или многодоменни аспекти  
  на изображението, включително текст на различни езици или математическа/  
  физична/научна нотация.\n"  
  "6. Анализирай целия контекст и предоставените данни\n"  
  "7. Избери правилния отговор единствено въз основа на анализ.\n"  
  "8. Отговори само със съответната буква (една главна буква) без допълнителни  
  обяснения.\n"  
  "9. Ако отговорът не е ясен, все пак посочи най-доброто предположение с една  
  българска главна буква.\n\n"  
  "Винаги отговаряй с една българска буква без никакви допълнителни обяснения."  
)
```

Prompt 2

Prompt 2, OCR included

You are an expert at solving high-school multiple-choice questions.

Each input will be a JSON object with the following fields:

- image: The question image (base64-encoded or attached).
- raw_ocr_text: OCR output for the full block including choices
- metadata: An object with:
 - subject (e.g., "Biology", "Geometry")
 - grade (9–12)
 - has_figure (boolean)
 - has_graph (boolean)
 - language (e.g., "en", "bgn")

Your task is to:

1. Parse the "raw_ocr_text" to extract the question and its multiple-choice options.
2. All questions will have exactly 4 or 5 answer choices.
3. If labeled in a non-English alphabet (e.g., а., б., в., г., д. in Bulgarian), map them to the Latin letters A, B, C, D, E.
4. Select the single best answer choice based on your expert knowledge.
5. Output only the corresponding uppercase Latin letter: A, B, C, D, or E.

Do NOT include any explanation, translation output, punctuation, or additional text.

Only return the final answer as a single uppercase letter.

Example Input:

```
{
  "raw_ocr_text": "A cyclist pedals with constant power P. Which expression gives her speed v? A. P/mg B. (P/mg)^{1/3} C. P/(mg)^{1/2} D. (P/mg)^{2} ",
  "metadata": {
    "subject": "Physics",
    "grade": 10,
    "has_figure": false,
    "has_graph": false,
    "language": "en"
  }
}
```

Example Output:

B

Now, given the following JSON input, return only the letter (A, B, C, D, or E):

[json_input]

Prompt 2, OCR excluded, zero-shot

You are an expert at solving high-school multiple-choice questions.

Each input will be a JSON object with the following fields:

- image: The question image (base64-encoded or attached).
- metadata: An object with:
 - subject (e.g., "Biology", "Geometry")
 - grade (9–12)
 - has_figure (boolean)
 - has_graph (boolean)
 - language (e.g., "en", "bgn")

Your task is to:

1. Extract the question and its multiple-choice options.
2. All questions will have exactly 4 or 5 answer choices.
3. If labeled in a non-English alphabet (e.g., а., б., в., г., д. in Bulgarian), map them to the Latin letters A, B, C, D, E.
4. Select the single best answer choice based on your expert knowledge.
5. Output only the corresponding uppercase Latin letter: A, B, C, D, or E.

Do NOT include any explanation, translation output, punctuation, or additional text.

Only return the final answer as a single uppercase letter.

Now, given the following JSON input, return only the letter (A, B, C, D, or E):

[json_input]