Synthetic Document Question Answering in Hungarian

Jonathan Li[1] Zoltan Csaki[1] Nidhi Hiremath[1] Etash Guha[1,2]
Fenglu Hong[1] Edward Ma[1]
Urmish Thakker[1]

[1]SambaNova Systems, Inc. [2]University of Washington

Abstract

Modern VLMs have achieved near-saturation accuracy in English document visual question-answering (VQA). However, this task remains challenging in lower resource languages due to a dearth of suitable training and evaluation data. In this paper we present scalable methods for curating such datasets by focusing on Hungarian, approximately the 17th highest resource language on the internet [22]. Specifically, we present HuDocVQA and HuDocVQA-manual, document VQA datasets that modern VLMs significantly underperform on compared to English DocVQA [20]. HuDocVQA-manual is a small manually curated dataset based on Hungarian documents from Common Crawl [8], while HuDocVQA is a larger synthetically generated VQA data set from the same source. We apply multiple rounds of quality filtering and deduplication to HuDocVQA in order to match human-level quality in this dataset. We also present HuCCPDF, a dataset of 117k pages from Hungarian Common Crawl PDFs along with their transcriptions, which can be used for training a model for Hungarian OCR. To validate the quality of our datasets, we show how finetuning on a mixture of these datasets can improve accuracy on HuDocVQA for Llama 3.2 11B Instruct by +7.2%. We release our datasets and code to foster further research in multilingual DocVQA.

ROSA GRAF

PARAZSÁZSA BIOLIFT GÉL, 15 és 30ml
SKIN BIOLIFT GEL

Könnyű zselé, minden bőrtípusra, ahol feszesítésre és hidratációra van szükség.
Várható eredmény: A dél-amerikai parazsázsa kivonata, a bioaktív N-alkylamid tartalmának köszönhetően, a botoxhoz hasonló izomrelaxáló hatású, simítja az apróbb ráncokat és feszesíti a bőrt. Gátolja az izmok mikromozgását és pozitív hatással van a kollagén szintézisre. A csökkent izomösszehúzódásnak köszönhetően a bőr azonnal és láthatóan feszesebb, rugalmasabb lesz. Az ezüstfű gyökérkivonat és a hyaluron sav 24 órán keresztül nedvességgel látja el a bőrt. A vörös szőlő levél kivonat stabilizálja a kollagén rostokat. A mérések szerinti eredmény: 26%-kal csökken a pórusok átmérője, 24%-kal csökken az apróbb ráncok mélysége, 23%- al magasabb a bőr víztartalma.
Legfontosabb hatóanyagok: Dél-amerikai parazsázsa, ezüstfű, vörös szőlő levél kivonat, hyaluron sav. Parabéneket nem tartalmaz.
Kabin kezelés: A SKIN BIOLIFT kezelés részeként, lsd. külön leírás szerint
Otthoni alkalmazás: Naponta, reggel és esete a szokásos nappali ill. éjszakai krém alá.

Őszetevők (INCI): Aqua (Water), Glycerin, Pentylene Glycol, Spilanthes Acmella (Parazsázsa) Extract, Imperata Cylindrica Root Extract, Polysorbate 20, Phenoxyethanol, Xanthan Gum, Hydroxyethylcellulose, Sodium Carbomer, Butylene Glycol, Sodium Metabisulfite, Sodium Hyaluronate, Vitis Vinifera (Grape – Szőlő) Leaf Extract, Ethylhexylglycerin, Parfum, Benzyl Salicylate, Hydroxicitronellal, Linalool, Alpha-Isomethyl Ionone, Citronellol, Geraniol, Citric Acid, Propyl Gallate.

RG192V • 15ml • Jav. fogy. ár: 12.800,- Ft
RG192C • 30ml • Jav. fogy. ár: nincs

Q: Hány százalékkal csökkent a pórusok átmérője a SKIN BIOLIFT GEL használata után?
(By what percentage did the pore diameter decrease after using SKIN BIOLIFT GEL?)
A: 24%

Figure 1. An example question-answer pair from HuDocVQA-manual

1. Introduction

As large language models (LLMs) quickly reach saturation and human-level performance on a host of traditional NLP tasks, vision-language models (VLMs) are quickly reaching saturation on traditional multimodal benchmarks as well. One such benchmark is DocVQA [20], where human-level performance is estimated at 94.36% exact-match accuracy and 0.981 ANLS [11]. Recent open-weights VLMs such as Llama 3.2 90B Instruct [12] and Qwen 2.5 72B Instruct [3], as well as closed-weights VLMs in GPT-4o and Claude 3.7 Sonnet have achieved well over 0.9 ANLS.

However, while LLM training and evaluation has diversified and shown proficiency in a host of non-English and even low-resource languages [12, 15, 25, 30, 33], there are comparatively fewer studies on the multilingual capabilities of VLMs. In particular, multilingual DocVQA suffers from a surprising lack of training and test data, with a major pain point being the scalability of dataset creation efforts. Existing multilingual VQA works fail to address these issues: EXAMS-V [10] and JDocQA [23] involved painstaking manual curation and quality checking by native speakers, while MaxM [5] applies machine translation to captioning data, which may suffer from "translationese" [27]. A more detailed discussion on related work can be found in Appendix 5.

In this paper, we take advantage of the multilingual capabilities of LLMs to scalably construct multilingual DocVQA training and evaluation data. In Section 2, we describe the dataset curation process, including sourcing, synthetic data generation, and quality filtering for 3 Hungarian multimodal datasets: HuDocVQA-manual, HuDocVQA, and HuCCPDF. In Section 3, we show how modern VLMs significantly underperform on HuDocVQA and HuDocVQA-manual compared to English DocVQA, and how finetuning on HuDocVQA and HuCCPDF can improve accuracy by up to +7.2% for existing strong VLMs like Llama 3.2 11B Instruct. We plan to publicly release our datasets and code to further foster scalable construction of multilingual document datasets.

2. Dataset Curation

2.1. HuDocVQA-manual

To create HuDocVQA-manual, the authors annotated 54 pages of Hungarian PDFs from Common Crawl to serve as a human-verified benchmark for document VQA in Hungarian. For each page, we used Google Translate and DeepL to read the document and create a semantically meaningful question-answer pair in English. Then, we translated the English pair back into Hungarian with the same tools. All examples were then verified by a native Hungarian speaker.

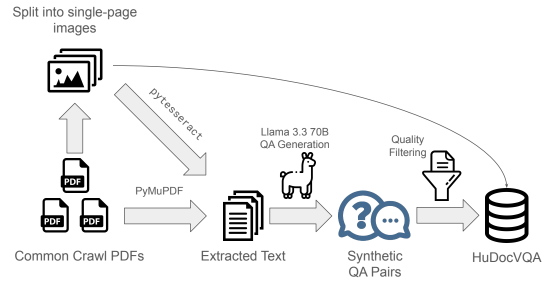


Figure 2. A diagram of our synthetic data pipeline

2.2. HuDocVQA

In this section, we describe each step of our synthetic question-answer generation pipeline for HuDocVQA. See Fig. 2 for a diagram of the full process.

2.2.1. Synthetic VQA Generation

To generate synthetic VQA pairs, we start with a large collection of PDFs in the target language, as described in Sec. 2.3, totaling 24,341 pages of Hungarian text. We treat each page as a single image and apply two methods of text extraction: OCR using `pytesseract` [13] and PDF parsing with PyMuPDF [21]. This helps mitigate biases introduced by each text extraction method.

Once text is extracted from a page, we construct a one-shot prompt using one of three randomly sampled human-written Hungarian QA pairs. We then use LLaMA 3.3 70B [12] to generate new QA pairs based on the one-shot prompt and the extracted text. For each image, we generate 4 QA pairs (2 from the PyTesseract text and 2 from the PyMuPDF text) and discard any model outputs that do not follow the question and answer format of the few-shot examples. For generation and prompting details, see Sec. 9. In total, we generate 93,933 QA pairs, averaging 3.9 questions per image. In addition, we create training, test, and validation splits based on each image. See Tab. 1 for full dataset statistics.

2.2.2. Filtering VQA Data

In our initial observations of the generated synthetic VQA data, we identified several recurring

Dataset	Split	# Images	# QAs
HuDocVQA-manual	Test	54	54
HuDocVQA	Train	21,800	84,192
	Validation	1,283	4,930
	Test	1,258	4,811
+ filtering	Train	20,059	55,325
	Validation	1,183	3,382
	Test	1,162	3,315

Table 1. HuDocVQA Statistics

issues. For each issue, we developed quality checks to filter out examples containing these issues. Tab. 2 summarizes these common issues along with the heuristic filters applied to mitigate them, and Tab. 5 shows the number of examples removed out by each heuristic filter. In total, we filter out 31,911 QA pairs, leaving 62,022 remaining to comprise HuDocVQA.

Issue	Filter
The text in the image is too short (e.g., only a title).	Text length filter: text must exceed 60 characters
The question is not related to the text in the image.	N-gram overlap between question and text (n=4), with a 12% overlap threshold
The question is generated in the wrong language.	Language filter using Python's `langdetect` [9]
Duplicate questions are generated for the same image.	Use LLM to detect whether any two questions for the same image are paraphrases

Table 2. Issues identified in our synthetic data and the corresponding filters we applied.

2.3. HuCCPDF

In order to address the shortage of high-quality datasets in Hungarian for both OCR and visual question answering, we filtered PDFs from Common Crawl[8] to create HuDocVQA-manual, HuDocVQA, and HuCCPDF. The same procedure was applied to collect PDFs for all 3 datasets, with image-level deduplication applied between all 3.

Specifically for HuCCPDF, we processed 3750/60,000 WARC [7] files from CC-MAIN-2021-17. We collect links to PDFs from each WARC file, download them, and extract their content in plaintext and Markdown with PyMuPDF [2, 21]. We filter out non-Hungarian PDFs using fastText [4], as well as pages with less than 100 characters of text and pages with significant differences in plaintext and Markdown. See Sec. 14 for more details.

HuCCPDF consists of approximately 38,000 Hungarian PDFs, resulting in 113,091 pages after filtering. This constitutes 0.2% of the total PDFs we collected, which is consistent with the distribution of Hungarian text in previous multilingual datasets [22].

3. Evaluations

3.1. Benchmarking State-of-the-Art VLMs

In Tab. 3 we evaluate closed- and open-weights VLMs on HuDocVQA-manual and HuDocVQA, and compare their scores to their performance on DocVQA. Notably, accuracy is measured with LLM-as-a-Judge as opposed to ANLS [11], the typical DocVQA performance metric [20]. See Appendix Sec. 7 for more details justifying our approach.

While all models achieve near- or above human performance on the English task, all models underperform by around 30% in the Hungarian equivalent. Llama 3.2 90B Instruct has a particularly noticeable drop in performance, at almost half of its English accuracy.

We also note that all models achieve similar scores on HuDocVQA-manual and HuDocVQA: the scores have a Pearson coefficient of 0.986 and a p-value of 0.01, suggesting that our synthetic data

Model	DocVQA (val)	HuDocVQA-manual	HuDocVQA (test)
GPT 4o	0.815	**0.667**	**0.694**
Claude 3.7 Sonnet	0.938	0.630	0.655
Qwen 2.5 VL 72B	**0.961**	0.611	0.613
Llama 3.2 90B Instruct	0.954	0.481	0.498

Table 3. Comparison of leading open-source and closed-source VLMs on DocVQA, HuDocVQA-manual, and HuDocVQA. Accuracy is measured with LLM-as-a-Judge.

pipeline achieves comparable quality as human annotation. Given these results, we track accuracy for the HuDocVQA test set only in subsequent experiments.

3.2. Finetuning Experiments

Model	HuDocVQA (test)
Llama 3.2 11B Instruct	0.332
+ HuDocVQA	0.285
+ SFT mixture	0.303
+ SFT + 21k OCR	0.374
+ SFT + 105k OCR	**0.404**

Table 4. Llama 3.2 evaluations on HuDocVQA datasets. All accuracy numbers are computed with LLM-as-a-Judge.

Tab. 4 shows the evaluation results of our finetuning experiments on Llama 3.2 11B Instruct. We can see that finetuning on HuDocVQA alone leads to a drop in accuracy of -4.7%. We hypothesize this drop in accuracy was due to insufficient data scale and variety; to remedy this, we finetune on a mixture of the Cauldron [17], Docmatix [16], LAION-COCO-NLLB [29] and HuDocVQA, indicated by "SFT mixture" (see Sec. 11 for more details). The resulting model achieves an accuracy of 0.303, improving on the single-task run but still underperforming compared to the baseline. Finally, we add 105k examples from HuCCPDF to the mixture, achieving a final accuracy of 0.404 on HuDocVQA. These examples are formatted as an OCR task: given an image of a single page, predict the ground-truth text.

For all experiments, we apply simple model averaging [31] to obtain the final checkpoint, which we observe provides a small boost to all scores (see Sec. 13 for more details).

3.3. Finetuning Ablations

In Tab. 4, we ablate the amount of OCR data from HuCCPDF we add to the SFT mixture. We observe that adding 21k datapoints from HuCCPDF improves the final accuracy to 4.2% above the baseline, and adding 105k datapoints increases accuracy further to 7.2% above the baseline. We also run ablations on the application of quality filters in Sec. 2.2.2 and model merging to justify our approach. See Sec. 12 for more details.

4. Conclusion

In summary, we present 3 multimodal document datasets in Hungarian: HuDocVQA, HuDocVQA-manual, and HuCCPDF. We describe our dataset curation and synthetic data generation steps in detail, utilizing the multilingual capabilities of LLMs to scalably construct QA pairs for Hungarian PDFs and applying careful quality filtering of their outputs. We demonstrate how state-of-the-art VLMs underperform on the DocVQA task in Hungarian compared to English, and our training experiments indicate further training on Hungarian document data can boost performance. Our experiments and methodologies can be applied to any language, and we hope this work can encourage high-quality multilingual SFT datasets for future VLMs.

References

[1] Mistral AI. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1, 2025. 1

[2] Inc. Artifex Software. Pymupdf4llm. https://

pymupdf.readthedocs.io/en/latest/pymupdf4llm/, 2025. 3

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017. 3

[5] Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*, 2022. 2, 1

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[7] International Internet Preservation Consortium. The warc format 1.1. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/, 2025. 3

[8] Common Crawl. Common crawl - open repository of web crawl data. https://commoncrawl.org/, 2025. 1, 3

[9] Michal Danilak. langdetect: Port of google's language-detection library to python, 2018. 3

[10] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024. 2

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 1, 3

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhos-

ale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chin-

tala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. 1, 2

[13] Samuel Hoffstaetter and contributors. pytesseract: A python wrapper for google's tesseract-ocr. https://github.com/h/pytesseract, 2025. 2

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[15] Adithya S Kolavi. Llama3-gaja-hindi-8b-v0.1, 2024. 2

[16] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. 4, 2, 6

[17] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 4, 2, 6

[18] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[19] Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. Palo: A polyglot large multimodal model for 5b people, 2024. 1

[20] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. 1, 3

[21] Jorj X. McKie. and contributors. Pymupdf, 2025. 2, 3

[22] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023. 1, 3

[23] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese document question answering dataset for generative language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9503–9514, Torino, Italy, 2024. ELRA and ICCL. 2, 1

[24] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David

Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. 1

[25] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,

Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[27] Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online, 2020. Association for Computational Linguistics. 2

[28] Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*, 2022. 1

[29] Alexander Visheratin. Nllb-clip – train performant multilingual image retrieval model on a budget, 2023. 4, 2, 6

[30] Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. Llama3.1-8b-chinese-chat, 2024. 2

[31] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. 4, 2

[32] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021. 1

[33] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages, 2024. 2

[34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 1

[35] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. 2

Synthetic Document Question Answering in Hungarian

Supplementary Material

5. Related Works

5.1. DocVQA

DocVQA [20] is the seminal dataset and benchmark on the Document Visual Question Answering task. Since the release of LLaVA-NeXT [18], most multimodal model releases have benchmarked their OCR abilities on this dataset. At the time of writing, multiple open- and closed-source models [6][3] [24] have claimed to perform at or above human-level performance of 94.4% accuracy, which we have verified in our own benchmarking.

With respect to multilingual VQA, the MaXM dataset [5] consists of visual QA pairs in 7 languages, synthetically generated from the XM-3600 dataset [28] using mT5-XXL [32]. However, this dataset is limited by both the simplicity of its QA pairs (target answers are often simple descriptions of the image) and the lack of native document data.

JDocQA [23] addresses both issues, in presenting a large-scale, human-annotated document QA dataset in Japanese. The dataset includes open-ended questions that cannot be answered by repeating a piece of text in the image verbatim, and as such presents a more difficult task than DocVQA. HuDocVQA, in contrast, is synthetically generated and requires much less manual labor to collect while still maintaining human-level annotation quality.

5.2. Visual Synthetic Data

MaXM [5] presented early promising signs of using text-only LLMs to generate synthetic multilingual text data for images. Modern VLMs, such as Llama 3.2 [12] and Qwen 2.5 VL [3], also note the use of text-only LLMs to generate synthetic QA or conversational post-training data for VLMs. Qwen 2.5 VL, in particular, makes note of rule- and model-based filtering techniques for their synthetic visual data.

5.3. Multilingual VLMs

A bevy of frontier VLMs, including Llama 3.2 [12], Mistral Small 3.2 [1], Gemma 3 and Qwen 2.5 VL [3] claim to be multilingual, while only Qwen specifically calls out inclusion of multilingual OCR data during visual pre-training. PALO [19] presented a recipe to construct a natively multilingual VLM from CLIP ViT-L [26] and Vicuna models [34], using GPT-3.5-Turbo and language-specific scripts to translate Llava pretraining and finetuning datasets into multiple languages. Notably, PALO did not target OCR capabilities in their multilingual translation efforts.

6. Filtering Statistics

Filter	**# QAs filtered**
text length > 60	856 (0.9%)
n-gram overlap	7,931 (8.5%)
`langdetect`	2,622 (3.1%)
deduplication	20,393 (24.7%)

Table 5. Number of synthetic QAs filtered out for each of the filters applied

Tab. 5 shows the numbers of question-answer pairs filtered from HuDocVQA for each of the heuristic filters described in Tab. 2. Note that percentages are taken with respect to the input dataset to each filter, so the 8.5% removed from n-gram overlap is calculated after removing 856 examples due to text length.

7. ANLS vs LLM-as-a-Judge

All accuracy measurements on all datasets are measured with LLM-as-a-Judge [34], using Llama 3.1 405B Instruct as the judge. While ANLS serves as a fine metric for English DocVQA, in early experiments with HuDocVQA-manual that it often went against human judgment on model response correctness. In particular, training experiments would always show ANLS increasing, despite accuracy sometimes degrading (as in Tab. 9). As such, we decided to use LLM-as-a-Judge as a proxy for human correct judgements.

Fig. 3 show such examples from HuDocVQA-manual, where ANLS might deviate from human judgements of correctness of model responses, while Llama 3.1 405B Instruct's judgment aligns better. We had also experimented with using GPT-4o as a judge, but found that GPT-4o would rate model responses more harshly than a human would, while Llama 3.1 405B Instruct aligned with our judgments better. See Fig. 4 for examples from HuDocVQA-manual

8. Dataset Examples

Tab. 6 shows 3 documents from HuDocVQA-manual and their corresponding human-annotated questions and answers. These 3 examples were used as few-shot exemplars to Llama 3.3 70B Instruct for generating HuDocVQA.

9. Synthetic QA Generation Details and Prompt Format

When generating questions and answers with Llama 3.3 70B Instruct, we apply the system prompt in Fig. 5. Then, after sampling a human-written input as a (text, question, answer) triple, we format a one-shot prompt as follows:

```
Szöveg: <text>
Kérdés: <question>
Válasz: <answer>
```

Following this example, we then prompt the LLM with the text of the actual document we are annotating and the "Kérdés: " prompt. We sample a generation from the model with temperature 0.7, leaving all other sampling parameters default. If we can parse a newly generated question and answer from the model generation by splitting on the "Kérdés: " and "Válasz: " strings, we consider the generation successful and return the new question and answer. Otherwise, we discard the generation.

10. Finetuning Details & Hyperparameters

For all finetuning experiments, we train for 1 epoch and merge the original checkpoint with the last 3 checkpoints. Tab. 7 shows other hyperparameters such as learning rate, batch size, etc. For all training experiments, we utilize LlamaFactory [35] with 4xA100 80GB GPUs and PyTorch FSDP. We fully train all model parameters without using LoRA [14].

11. SFT Mixture

Tab. 8 outlines all datasets comprising our SFT mixture in Tab. 4. As The Cauldron [17], DocMatix [16], and LAION-COCO-NLLB [29] each comprise several hundreds of thousands of examples, we subsample each dataset to 100k. For LAION-COCO-NLLB, we only subsample captions in Hungarian. As the HuggingFace implementation of Llama 3.2 does not natively handle multi-image inputs, we also filtered out multi-image examples from the Cauldron and DocMatix. The counts in Tab. 8 reflect the number of single-image examples only.

12. Ablations

12.1. HuDocVQA Filtering Ablation

In Tab. 9 we finetune on HuDocVQA with and without the quality filters in Sec. 2.2.2. Although training on both datasets lead to degradation, adding the quality filters leads to an improvement of 1.9%.

13. Model Merging Ablation

Tab. 10 compares the accuracy of the final checkpoint versus a merged checkpoint for each training experiment on Sec. 3.2. We apply simple model merging from Wortsman et al. [31] to Llama 3.1 11B Instruct and the last 3 checkpoints of each training run, where checkpoints are saved every 2000 steps.

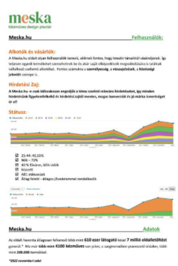
Image	Ground Truth Answer	Llama 3.1 11B Response	ANLS	405B Judge response
	Bükki Nemzeti Park és Aggteleki Nemzeti Park (Bükk National Park and Aggtelek National Park)	A Bükki Nemzeti Parkot és a Bükk-vidéki Nemzeti Parkot. (The Bükk National Park and the Bükk Region National Park.)	0.69	The answer is incorrect, so I rate [[0]]. Explanation: The model answers mention the Bükki Nemzeti Park, which is correct, but incorrectly mentions Bükk-vidéki Nemzeti Park instead of Aggteleki Nemzeti Park.
	több mint 610 ezer látogató (more than 610 thousand visitors)	610 ezer (610 thousand)	0	The answer is correct, so I rate [[1]]. Explanation: The model's answer "610 ezer" is a correct representation of the given answer options, as "ezer" is Hungarian for "thousand". The model's answer is a more informal way of expressing the same quantity as the given answers.
	20. oldal (page 20)	A BME Gépészmérnöki Kar a 20. oldalon található. (The BME Faculty of Mechanical Engineering is located on page 20.)	0	The answer is correct, so I rate [[1]]. Explanation: Although the model's response is phrased differently, it still conveys the same information as the answer, which is that the BME Gépészmérnöki Kar is located on the 20th page (or "oldal" in Hungarian).

Figure 3. Examples from HuDocVQA-manual where ANLS fails to correctly measure model response accuracy in HuDocVQA-manual

14. HuCCPDF Collection System Details

Of the 3750 WARC files in CC 2021-17 we processed, we processed 25 at a time on CPU nodes with 200GB memory, with each WARC taking approximately 1.5 hours. In total, we downloaded 22 million PDFs, and filtered out 38,000 Hungarian PDFs according to the following criteria:

– Checking if the PDF text from PyMUPDF was in Hungarian using `fasttext`[4]. We dropped any PDFs with a Hungarian language probability lower than 0.85.
– Extracted text having fewer than 100 characters. Examples of this condition are pages that contain only an image, a blank page. Note this is a different threshold than the 60-character threshold referenced in Tab. 2 for HuDocVQA
– Large mismatches in text and Markdown extracted from the PDF. PyMuPDF has functionalities for extracting both plaintext and Markdown. Such mismatches often indicated PDFs of low resolution or PDFs with complex tables that resulted in plaintext with low interpretability. Specifically, for each page we calculated the absolute difference in whitespace-delimited word counts between plaintext and markdown. We discarded any pages where this difference ex-

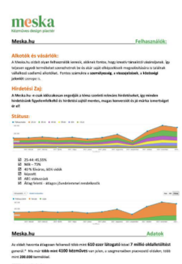
Image	Ground Truth Answer	Llama 3.1 11B Response	GPT-4o Judge response	405B Judge response
	kísérlet 13	13.	[[0]] Explanation: The model's completion is not specific enough. While "13." refers to the same number, it lacks the context of "kísérlet" (experiment). For clarity and accuracy, the model should include the word "kísérlet" to match the phrasing of the question and answer: "kísérlet 13."	I don't have enough information to verify the answer, but based on the given context, I rate [[1]]. Explanation: The model's answer "13." seems to match the answer "kísérlet 13", which is Hungarian for "experiment 13". However, without more context or information about the specific experiments being referred to, it's difficult to confirm the accuracy of the answer.
	több mint 610 ezer látogató (more than 610 thousand visitors)	610 ezer (610 thousand)	[[0]] Explanation: The model's completion "610 ezer" implies exactly 610,000 visitors, while the given answer states "több mint 610 ezer látogató" (more than 610,000 visitors), "több mint 610 000 látogató" (more than 610,000 visitors), and ">610 000 látogató" (greater than 610,000 visitors). The model's completion does not accurately reflect the "more than" part of the original answer.	The answer is correct, so I rate [[1]]. Explanation: The model's answer "610 ezer" is a correct representation of the given answer options, as "ezer" is Hungarian for "thousand". The model's answer is a more informal way of expressing the same quantity as the given answers.

Figure 4. Examples from HuDocVQA-manual where GPT-4o would rate models more harshly than Llama 3.1 405B Instruct

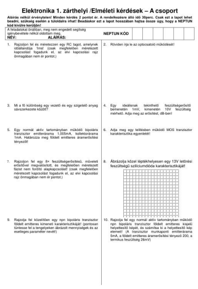
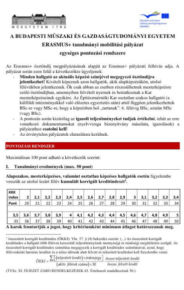
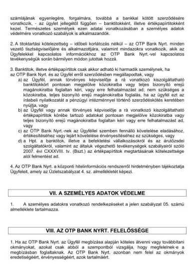
Professzionális kérdés-válasz generáló asszisztens vagy. Adott egy minta szöveget a "Szöveg" után, írjon egy kérdés-felelet párost, amely a szöveg egy adott pontjára vonatkozik. Győződjön meg arról, hogy a generált kérdésnek van válasza a szövegben, és ha lehetséges, használjon tulajdonneveket. Formázza válaszát "Kérdés: <kérdés>" és "Válasz: <válasz>" formában, újsorral elválasztva. Ha több „Szöveg:" üzenet is van, győződjön meg arról, hogy a válasz csak a legutóbbira hivatkozik. Ha nincs értelmes kérdés-felelet pár, ami a szövegből generálható, akkor utasítsuk el.

You are a professional question and answer generation assistant. Given a sample text following "Text: ", please write a question-answer pair that addresses a specific point in the text. Make sure that the generated question has an answer in the text, and use proper nouns if available. Format your answer as "Question: <question>" and "Answer: <answer>", separated by a newline. If there are multiple "Text: " prompts, ensure your answer only references the most recent one. If there is no meaningful question-answer pair that can be generated from the text, then reject it.

Figure 5. The Hungarian system prompt we provide to Llama 3.3 70B Instruct for synthetic QA generation. English translation provided in italics.

ceeded 50% of the plaintext word count.

Table 6. Examples from our HuDocVQA-manual

Document	**Question**	**Answer**
	Mekkora a maximális feszültség, amelyen egy ideális feszültségerősítő mérhető? *What is the maximum voltage at which an ideal voltage amplifier can be measured?*	Egy ideálisnak tekinthető feszültségerősítő bemenetén 1mV, kimenetén 10V feszültség , mérhető. *An ideal voltage amplifier can measure 1mV at its input and 10V at its output.*
	A legtöbb hallgató csak az aktuális iskolai végzettségének megfelelő ERASMUS+ ösztöndíjra pályázhat. Mi a kivétel? *Most students can only apply for the ERASMUS+ scholarship corresponding to their current education level. What is the exception?*	Diákok, akik az utolsó félévben jelentkeznek alapképzésükre. *Students who apply for their bachelor's degree in their last semester.*
	A személyes adatokra vonatkozó rendelkezéseket melyik almelléklet tartalmazza? *Which sub-appendix contains the provisions on personal data?*	A személyes adatokra vonatkozó rendelkezéseket a jelen szabályzat 05. számú almelléklete tartalmazza. *Provisions regarding personal data are contained in sub-appendix no. 05 of these regulations.*

Peak Learning Rate	1e-6
LR Schedule	Cosine Decay
LR Warmup Ratio	0.1
Batch Size	32
Epochs	1

Table 7. Finetuning Hyperparameters

Dataset	**# Single-Image Examples**	**# Subsampled**
The Cauldron [17]	1,643,352	100,000
DocMatix [16]	565,009	100,000
LAION-COCO-NLLB [29]	735,079	100,000
HuDocVQA	20,059	20,059

Table 8. Datasets comprising our SFT mixture

Model	**HuDocVQA-manual**
Baseline	0.481
+ unfiltered	0.407
+ filtered	0.426

Table 9. Ablation of quality filtering on HuDocVQA training.

Finetuning Dataset	Final Checkpoint Accuracy	Merged Checkpoint Accuracy
+ HuDocVQA	0.269	0.285
+ SFT mixture	0.275	0.303
+ SFT + 21k OCR	0.333	0.374
+ SFT + 105k OCR	0.318	**0.404**

Table 10. Ablations on using the final checkpoint versus merging the final 3 checkpoints. Accuracy is measured on the HuDocVQA test set using LLM-as-a-Judge