olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models

Anonymous Authors[1]

Abstract

PDF documents offer trillions of novel, high-quality tokens for language model training, but their diverse formats and layouts complicate content extraction. Traditional open source tools yield lower quality results than vision language models (VLMs), yet the best VLMs are costly (e.g., over $6,240 per million PDF pages for GPT-4o) or inaccessible when working with proprietary documents. We present olmOCR, an open-source toolkit for converting PDFs into clean, linearized plain text in natural reading order while preserving structure such as sections, tables, and equations. Our toolkit uses a fine-tuned 7B VLM trained on 260,000 pages from over 100,000 varied PDFs, including graphics, handwritten text, and poor scans. olmOCR is optimized for large-scale batch processing, converting a million pages for only $176. We find olmOCR outperforms even top VLMs including GPT-4o, Gemini Flash 2 and Qwen-2.5-VL on olmOCR-Bench, a curated set of 1,400 challenging PDFs with fine-grained unit tests that remain challenging even for the best tools and VLMs. We openly release all components of olmOCR: our fine-tuned VLM model, training code and data, an efficient inference pipeline that supports vLLM and SGLang backends, and benchmark.[1]

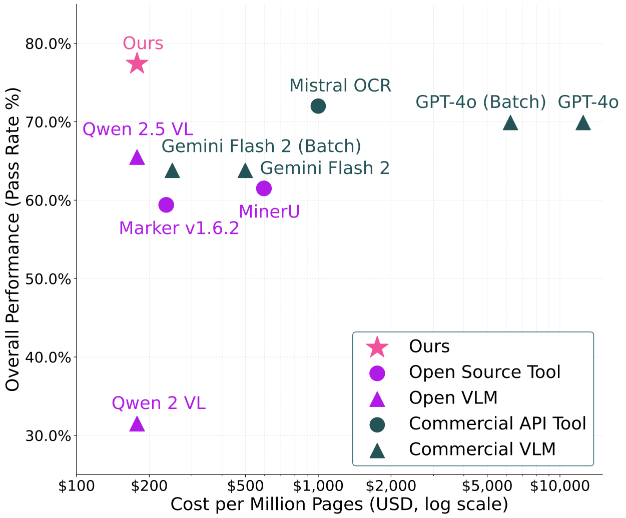


Figure 1. Performance-to-cost of olmOCR vs other tools or models for PDF linearization and content extraction.

1. Introduction

Access to clean, coherent text is essential for training modern language models (LMs) on trillions of tokens from billions of documents (Soldaini et al., 2024a; Penedo et al., 2024b; Li et al., 2024a); noisy or low-fidelity data can cause training instabilities and harm downstream performance (Penedo et al., 2023b; Li et al., 2024a; OLMo et al., 2024). Electronic documents, particularly PDFs, represent a significant repository of textual content, with trillions of documents stored in this format (PDF Association staff, 2015), which makes them critical for the development of language models For example, Qwen 3 (Yang et al., 2025) described training on "trillions of tokens" from PDFs.

Faithful extraction and representation of digitized print documents has been studied since the 1950s, with commercial OCR tools emerging in the late 1970s (Mori et al., 1992). Tesseract's 2006 release was a major milestone as a high-quality, open-source OCR toolkit (Smith, 2013). Modern PDF extraction tools are either **pipeline-based systems**—comprising multiple ML components (e.g., MinerU (Wang et al., 2024a), Marker (Paruchuri, 2025), Grobid (gro, 2008–2025), VILA (Shen et al., 2022), PaperMage (Lo et al., 2023a))—or **end-to-end models**, which parse documents in a single step (e.g., Nougat (Blecher et al., 2023), GOT Theory 2.0 (Wei et al., 2024)). While pipeline-based systems emphasize faithful extraction, end-to-end models have advanced **linearization**, addressing the challenge of preserving logical reading order in complex layouts. Recent proprietary VLMs have significantly improved end-to-end linearization and extraction (Bai et al.,

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Code is made anonymous for review at https://anonymous.4open.science/r/olmocr-F583/.

2025; Google, 2025), but at high cost—for example, processing a million pages with GPT-4o can exceed $6,200 USD.[2]

We introduce **OLMOCR**, a general-purpose context extraction and linearization toolkit to convert PDFs or images of documents into clean plain text suitable for language model development. Our contributions in this work are as follows:

- **Data.** We create olmOCR-Mix, a collection of 260,000 crawled PDF pages paired with their OCR output by GPT-4o, that we use to train our models. These documents represent a diverse set of publicly available PDFs, with a skew towards academic papers, public domain books, legal documents, brochures, and more.
- **Benchmark.** We develop OLMOCR-BENCH, a comprehensive benchmark for evaluating document extraction tools. The benchmark covers 1,400 PDF pages with over 7,000 unit-test cases spanning diverse document types.
- **Model and Code.** We fine-tune Qwen2-VL-7B-Instruct (Wang et al., 2024b) on olmOCR-Mix, producing olmOCR-7B. We package our VLM in the OLMOCR Python toolkit, written to scale efficiently from one to hundreds of GPUs using SGLang (Zheng et al., 2024) inference engine. OLMOCR achieves state-of-the-art performance on our benchmark, even outperforming Qwen-2.5-VL-7B while remaining more cost-effective than existing alternatives, including commercial APIs; OLMOCR can produce high-quality plain text at less than $176 per million PDF pages.

2. Creating and Training on olmOCR-Mix

2.1. Crawling PDFs

We randomly sample PDFs from an internal dataset of 240 million PDFs crawled from public internet sites, as well as PDFs of public domain books sourced from the Internet Archive. While the web crawled set is often born-digital documents, PDFs from the Internet Archive consist of image scans. We then perform a set of filters: Using the Lingua package (Emond, 2025), we identify and filter out non-English documents. Further, we remove any document that failed to be parsed by pypdf, contains spam keywords, is a fillable form, or whose text is too short. We then sampled (up to) three pages uniformly at random from each PDF. Our final set consists of 96,929 unique web-crawled PDF documents (totaling 240,940 pages) and 5,896 Internet Archive books (totaling 17,701 pages), for an overall total of 102,825 documents and 258,641 pages. Among the PDFs in the training set, 55.9% are academic documents, 11.2% are brochures, 10.2% are legal documents, 6.8% are books, 5.6% are tables, 4.7% are diagrams, 1.9% are slideshows, and 3.7% fall into other categories.

2.2. Generating Linearized Plain Text

We generated supervision data for PDF-to-plain-text conversion using GPT-4o, as human annotation is expensive and existing PDF extraction tools are unreliable, especially for document images.[3] To address the model's occasional omissions and hallucinations—particularly on complex layouts—we augmented the PDF page images with extracted text blocks and layout information, using a noisy but useful PDF internal representation from pypdf (PyPDF, 2012–2025). We prompted GPT-4o with this combined visual and text input (called DOCUMENT-ANCHORING) to produce our final supervision targets. See Appendix Figure 2 for example and §B.1 for prompt.

2.3. Fine-Tuning olmOCR-7B

Starting from a Qwen2-VL-7B-Instruct checkpoint, we fine-tune olmOCR-7B on olmOCR-Mix. Training is implemented using Hugging Face's transformers library (Wolf et al., 2020). We use an effective batch size of 4, learning rate of 1e-6, AdamW optimizer, and a cosine annealing schedule for 10,000 steps (roughly 1.2 epochs). We use single node with 8 x NVIDIA H100 (80GB) GPUs. A single training run took 16 node hours, with all training experiments totaling 365 node hours.

During fine-tuning, we slightly alter the DOCUMENT-ANCHORING prompt, removing some instructions and shrinking the image size so that PDF pages are rendered to a maximum dimension of 1024 pixels on the longest edge. The simplified text prompt is in Appendix §B.3. Loss was masked so only the final response tokens participated in the loss calculation.

3. Building OLMOCR-BENCH

We develop OLMOCR-BENCH to systematically evaluate PDF linearization and content extraction performance across diverse tools and models. OLMOCR-BENCH operates by assessing a series of predefined pass-or-fail "unit-tests"—*Given an input whole PDF, does the plain text output satisfy a specific property or contain a specific element?* Each test is designed to be simple, unambiguous, and deterministically machine-verifiable. OLMOCR-BENCH comprises

[2] Batch pricing at $1.25 USD (input) and $5.00 USD (output) per 1M tokens in Feb 2025. Details in Appendix§A.

[3] In October 2024, we evaluated several leading VLMs for data generation. Gemini 1.5 was eliminated due to frequent RECITATION errors (though this was resolved by February 2025), GPT-4o mini produced excessive hallucinations, and Claude Sonnet 3.5 was cost-prohibitive. We selected gpt-4o-2024-08-06 as it offered the optimal balance of accuracy, reliability, and cost-efficiency in batch mode.

1,402 distinct PDF documents derived from diverse source repositories, covered by 7,010 unique test cases.

3.1. Unit Test Categories

We designed five distinct test categories to assess different aspects of linearization and context extraction performance. **Text Presence** checks that a specific text segment appears in the plain text output, with options for fuzzy matching and positional constraints. In contrast, **Text Absence** ensures that a given segment does *not* appear—useful for filtering out headers, footers, or pagination. The **Natural Reading Order** category validates that two text segments appear in the correct order, while allowing for some flexibility and fuzzy matching. **Table Accuracy** evaluates whether a table cell and its neighbors in the output match expected values, supporting both Markdown and HTML formats. **Math Formula Accuracy** involves verifying that a math equation is present by comparing the visual layout of symbols to a rendered reference. Finally, the **Baseline** category confirms that the output includes reasonable alphanumeric text and avoids common issues like repeating patterns or unwanted character sets.

3.2. Sourcing Documents and Creating Tests

We define seven document types that posed challenges for olmOCR and developed custom acquisition strategies for each. We filtered out documents containing PII and were not meant for public dissemination and performed URL-level deduplication against olmOCR-Mix (Soldaini et al., 2024a). We created test cases using a mix of manual design and GPT-4o prompting; see Appendix §C for details and examples.

Our dataset construction drew from a range of sources and document types. The **arXiv Math (AR)** dataset consists of recent arXiv math papers with single TeX source files; for these, we identified and validated LaTeX expressions using our pipeline and manual review. The **Old Scans Math (OSM)** dataset was built by extracting pages with formulas from old public domain math textbooks, with each formula manually annotated as a test case. For the **Tables (TA)** dataset, we sampled PDFs containing tables, used Gemini-Flash-2.0 to generate cell relationship tests, and then manually reviewed the results. The **Old Scans (OS)** dataset comprises historical letters and typewritten documents with transcriptions from the Library of Congress[4] digital archives; we generated Natural Reading Order test cases for these and manually checked them for accuracy. In constructing the **Headers Footers (HF)** dataset, we sampled additional documents from our internal crawled PDFs, identified header and footer regions using DocLayout-YOLO (Zhao et al., 2024), extracted their content with Gemini-Flash-2.0, and manually reviewed to ensure that such text is excluded from the linearized output. The **Multi Column (MC)** dataset includes multi-column PDFs sampled from our internal collection; for these, we used Claude-Sonnet-3.7 to extract the text order and manually verified that the text blocks were simple and coherent. Finally, the **Long Tiny Text (LTT)** dataset was created by crawling densely printed pages from the Internet Archive, generating test cases using Gemini-Flash-2.0, and manually verifying them.

3.3. Scoring

We run each of the PDF pages across each of our tools and methods to produce a markdown or plain text document. As all tests are Pass/Fail, we simply report percentage of tests passed, macro-averaged by document type.

4. Evaluating olmOCR

4.1. olmOCR-Bench Results

Table 1 shows evaluation results of olmOCR on olmOCR-Bench against a range of linearization tools and VLMs. We see that olmOCR significantly outperforms both the best commercial dedicated OCR tool (Mistral) as well as both GPT-4o, its teacher model, and Qwen 2.5 VL, which is an update to Qwen 2 VL, which was the base model for olmOCR-7B. We note that we developed olmOCR-Bench *after* training olmOCR-7B to prevent unfairly iterating on the benchmark before comparing with other methods. Qualitatively, olmOCR produces significantly cleaner plain text than specialized open-source tools (visualized in Appendix §E).

4.2. Downstream Evaluation

We demonstrate value of olmOCR for curating language model pretraining data. Following (Blakeney et al., 2024; Grattafiori et al., 2024; OLMo et al., 2024), we experiment with continued pretraining of OLMo-2-1124-7B (OLMo et al., 2024) using content extracted from a fixed collection of PDFs but ablating the use of olmOCR. For our baseline, we use tokens from peS2o (Soldaini & Lo, 2023), academic papers derived using Grobid (gro, 2008–2025) from the S2ORC (Lo et al., 2020) paper collection and further cleaned with heuristics for language modeling. Switching to using olmOCR processing results in a **+1.3 percentage point average improvement** on widely-reported LM benchmark tasks.[5].

[4] https://crowd.loc.gov

[5] Average of 55.2 (baseline) vs 53.9 (ours) over tasks including MMLU (Hendrycks et al., 2021), ARC$_C$ (Clark et al., 2018), DROP (Dua et al., 2019), HellaSwag (Zellers et al., 2019), NaturalQuestions (Kwiatkowski et al., 2019), WinoGrande (Sakaguchi et al., 2019).

Table 1. Evaluation results on OLMOCR-BENCH grouped by document types. Best unit test pass rate in each column is bold. 95% CI calculated by bootstrapping with 10k samples. Costs for API models using batch mode and for open VLM based on NVIDIA L40S.

Model	AR	OSM	TA	OS	HF	MC	LTT	Base	Overall	Cost per 1M pages
GOT OCR	52.7	52.0	0.2	22.1	93.6	42.0	29.9	94.0	48.3 ± 1.1	—
Marker v1.6.2	24.3	22.1	69.8	24.3	87.1	71.0	76.9	**99.5**	59.4 ± 1.1	$235
MinerU v1.3.10	75.4	47.4	60.9	17.3	**96.6**	59.0	39.1	96.6	61.5 ± 1.1	$596
Mistral OCR API	**77.2**	67.5	60.6	29.3	93.6	71.3	77.1	99.4	72.0 ± 1.1	$1,000
GPT-4o	51.5	**75.5**	69.1	40.9	94.2	68.9	54.1	96.7	68.9 ± 1.1	$6,240
Gemini Flash 2	32.1	56.3	61.4	27.8	48.0	58.7	**84.4**	94.0	57.8 ± 1.1	$249
Qwen 2 VL	19.7	31.7	24.2	17.1	88.9	8.3	6.8	55.5	31.5 ± 0.9	$176
Qwen 2.5 VL	63.1	65.7	67.3	38.6	73.6	68.3	49.1	98.3	65.5 ± 1.2	$176
Ours	75.6	75.1	**70.2**	**44.5**	93.4	**79.4**	81.7	99.0	**77.4 ± 1.0**	$176

5. Deploying OLMOCR

When considering real-world use, cost efficiency is just as important as performance.

Inference Pipeline. We deploy OLMOCR using SGLang (Zheng et al., 2024) for large-scale document processing. Documents are batched (∼ 500 pages each) and processed on GPU workers, scaling easily from single to hundreds of nodes via a shared cloud bucket (e.g., S3). Workers queue and process all PDF pages in a batch together, maximizing GPU utilization and throughput.

As shown in Table 1, OLMOCR is significantly cheaper than both API and other local models—over 32× cheaper than GPT-4o and 6× cheaper than MinerU. To contextualize the value of OLMOCR, at 1,000 tokens per page, to process all of peS2o PDFs can already cost $10.3M in H100 usage. In comparison, Mistral OCR is a commercial API tool specializing in this task, yet is over five times more expensive, making it even more prohibitive to use for language modeling. See Appendix §A for details on pricing and cost calculations.

Improving Robustness. Benchmark performance alone doesn't guarantee real-world usability, so we employ several additional techniques to ensure reliability. For **Prompt Format**, we make sure that prompts match the training format, and if the length exceeds 8,192 tokens, we simply shorten the DOCUMENT-ANCHORING tokens until everything fits. With **Retries**, we rely on the model's fine-tuning to keep outputs structured, so we don't require strict schema enforcement—if a JSON parse fails, we just try again. When it comes to **Rotations**, any pages flagged for rotation are automatically corrected and reprocessed. For **Decoding**, we watch for output repetitions and, if they occur, retry with a higher generation temperature and different anchor tokens; if problems persist, we fall back on text extraction. Further optimizations to abort failed generations earlier are planned for future work.

6. Related Work

PDF Linearization. Many tools exist for linearizing PDFs to plain text, ranging from basic parsers and OCR to advanced models like LayoutLM (Xu et al., 2020), VILA (Lin et al., 2024), and production systems such as PaperMage (Lo et al., 2023b), Grobid (gro, 2008–2025), but comprehensive VLM-based libraries for this task remain scarce, a gap our work addresses while comparing to recent models like Mistral (Mistral, 2025) and Qwen VL (Bai et al., 2023).

Benchmarking VLMs on Linearization. Existing benchmarks for document linearization, like FUNSD (Guillaume Jaume, 2019), SROIE (Huang et al., 2019), and RVL-CDIP (Harley et al., 2015), are domain-limited and task-specific, whereas our approach introduces a broader, unit-test-style evaluation spanning diverse document types and extraction tasks (e.g., tables (Zhong et al., 2020), formulas (Zhong et al., 2021)) and supports flexible tokenization.

Linearization for Language Modeling. While there is significant research on data curation for language modeling (Soldaini et al., 2024b; Penedo et al., 2024a; Li et al., 2024b; Wettig et al., 2025; Liu et al., 2024), little attention has been given to how linearization quality affects downstream model training, especially for PDF content—a gap this work seeks to fill, unlike prior efforts focused on web content (e.g., DCLM (Li et al., 2024b), RefinedWeb (Penedo et al., 2023a), OpenWebMath (Paster et al., 2023)).

7. Conclusion

We present OLMOCR, an open-source toolkit that efficiently converts PDFs to clean text, matching commercial performance at lower cost. We release our model, training set (`olmOCR-Mix`), and a comprehensive benchmark (OLMOCR-BENCH) of 7,010 unit tests across 1,403 PDFs. We hope OLMOCR will unlock new training sources of high-quality PDF documents that are currently underrepresented amid heavy reliance on crawled web pages.

References

Grobid. https://github.com/kermitt2/grobid, 2008–2025.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-VL technical report. *arXiv [cs.CV]*, February 2025.

Blakeney, C., Paul, M., Larsen, B. W., Owen, S., and Frankle, J. Does your data spark joy? performance gains from domain upsampling at the end of training, 2024. URL https://arxiv.org/abs/2406.03476.

Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. Nougat: Neural optical understanding for academic documents, 2023. URL https://arxiv.org/abs/2308.13418.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR*, arXiv:1803.05457, 2018.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 2019. URL https://arxiv.org/abs/1903.00161.

Emond, P. Lingua-py: Natural language detection for python, 2025. URL https://github.com/pemistahl/lingua-py. Accessed: 2025-01-06.

Google. Explore document processing capabilities with the gemini API. https://web.archive.org/web/20250224064040/https://ai.google.dev/gemini-api/docs/document-processing?lang=python, 2025. Accessed: 2025-2-23.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathurx, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guillaume Jaume, Hazim Kemal Ekenel, J.-P. T. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*, 2019.

Harley, A. W., Ufkes, A., and Derpanis, K. G. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., and Jawahar, C. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A., Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldaini, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. DataComp-LM: In search of the next generation of training sets for language models. *arXiv [cs.LG]*, June 2024a.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S. Y., Bansal, H., Guha, E., Keh, S. S., Arora, K., et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024b.

Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models, 2024. URL https://arxiv.org/abs/2312.07533.

Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., and Lin, M. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Lo, K., Shen, Z., Newman, B., Chang, J., Authur, R., Bransom, E., Candra, S., Chandrasekhar, Y., Huff, R., Kuehl, B., Singh, A., Wilhelm, C., Zamarron, A., Hearst, M. A., Weld, D., Downey, D., and Soldaini, L. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In Feng, Y. and Lefever, E. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 495–507, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.45. URL https://aclanthology.org/2023.emnlp-demo.45/.

Lo, K., Shen, Z., Newman, B., Chang, J. Z., Authur, R., Bransom, E., Candra, S., Chandrasekhar, Y., Huff, R., Kuehl, B., et al. Papermage: a unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 495–507, 2023b.

Mistral. Mistral ocr, 2025. URL https://mistral.ai/news/mistral-ocr.

Mori, S., Suen, C. Y., and Yamamoto, K. Historical review of OCR research and development. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, 80(7):1029–1058, July 1992. ISSN 0018-9219,1558-2256. doi: 10.1109/5.156468.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Guerquin, M., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J. D., Murray, T. C., Nam, C., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L. S., Farhadi, A., Smith, N. A., and Hajishirzi, H. 2 OLMo 2 Furious. *arXiv preprint*, 2024. URL https://api.semanticscholar.org/CorpusID:275213098.

Paruchuri, V. Marker: Convert pdf to markdown + json quickly with high accuracy, 2025. URL https://github.com/VikParuchuri/marker. Version 1.4.0.

Paster, K., Santos, M. D., Azerbayev, Z., and Ba, J. Openwebmath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*, 2023.

PDF Association staff. Pdf in 2016: Broader, deeper, richer. *PDF Association*, December 2015. URL https://pdfa.org/pdf-in-2016-broader-deeper-richer/.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023a.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R.-A., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116, 2023b. URL https://api.semanticscholar.org/CorpusID:259063761.

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C. A., Von Werra, L., Wolf, T., et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024a.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale, 2024b. URL https://arxiv.org/abs/2406.17557.

PyPDF. Pypdf: A pure-python pdf library. https://github.com/py-pdf/pypdf, 2012–2025. Accessed: 2025-01-06.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Shen, Z., Lo, K., Wang, L. L., Kuehl, B., Weld, D. S., and Downey, D. VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *Transactions of the Association for Computational Linguistics*, 10:376–392, 2022. doi: 10.

1162/tacl_a_00466. URL https://aclanthology.org/2022.tacl-1.22/.

Smith, R. W. History of the tesseract OCR engine: what worked and what didn't. In Zanibbi, R. and Coüasnon, B. (eds.), *Document Recognition and Retrieval XX*, volume 8658, pp. 865802. SPIE, February 2013. doi: 10.1117/12.2010051.

Soldaini, L. and Lo, K. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, https://github.com/allenai/pes2o.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A. H., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, P., Zettlemoyer, L., Smith, N. A., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024a. URL https://arxiv.org/abs/2402.00159.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A. H., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, P., Zettlemoyer, L., Smith, N. A., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024b. URL https://arxiv.org/abs/2402.00159.

Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., Zhang, B., Wei, L., Sui, Z., Li, W., Shi, B., Qiao, Y., Lin, D., and He, C. Mineru: An open-source solution for precise document content extraction, 2024a. URL https://arxiv.org/abs/2409.18839.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

Wettig, A., Lo, K., Min, S., Hajishirzi, H., Chen, D., and Soldaini, L. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, KDD '20, pp. 1192–1200. ACM, August 2020. doi: 10.1145/3394486.3403172. URL http://dx.doi.org/10.1145/3394486.3403172.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Zhao, Z., Kang, H., Wang, B., and He, C. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024.

Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., Barrett, C., and Sheng, Y. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

Zhong, X., ShafieiBavani, E., and Jimeno Yepes, A. Image-based table recognition: data, model, and evaluation. In

European conference on computer vision, pp. 564–580. Springer, 2020.

Zhong, Y., Qi, X., Li, S., Gu, D., Chen, Y., Ning, P., and Xiao, R. 1st place solution for icdar 2021 competition on mathematical formula detection, 2021. URL `https://arxiv.org/abs/2107.05534`.

A. Cost Estimates of PDF Extraction Systems

To estimate prices (Table C.1.6), we use rates provided by RunPod[6] as of February 2025. It prices a single on-demand NVIDIA L40S GPU at $0.79 USD per hour, and NVIDIA H100 80GB SXM at $2.69 USD per hour. Using these rates, costs (in USD) were computed as follows:

- **GPT-4o**: We evaluated GPT-4o in February 2025. We tested 1288 pages, which resulted in 3,093,315 input tokens at 833,599 output tokens. Priced at $2.50 per million input tokens and $10.00 per million output tokens, it resulted in a total of $16.07. Batch processing is priced at half of the cost, $8.03.
- **Mistral OCR**: As of May 2025, Mistral prices their OCR service at $1 per 1,000 pages, regardless of number of generated tokens.
- **MinerU**: We run the toolkit (version 1.3.10) on a single NVIDIA L40S GPU. It processed 1,288 pages in 58 minutes 22 seconds, costing $0.767.
- **Marker**: We run marker locally on L40S, version 1.6.2. In our test on 1,166 pages, it took 20 minutes, 52 seconds to parse 1166; we consider this second run and estimate its cost to $0.274. We note that this is much more cost-effective than using Marker APIs, which are priced at $1.5 per 1000 standard pages, and $3.0 per 1000 pages with layout/tables.
- **Gemini Flash 2.0**: As of February 2025, it is priced $0.10 per 1 million input tokens, and $0.40 per 1 million output tokens. In our testing over the same 1,288 pages used to evaluate GPT-4o, it cost in $0.643.
- **OLMOCR**: We tested OLMOCR on both L40S and H100 GPUs. On L40s, it processed 1,288 test pages in 17 minutes, 10 seconds. The effective throughput of the model was 906 output tokens per second, plus a 12% reties rate. Overall, we estimate its costs at $0.226. On H100, OLMOCR generates 3,050 output tokens per second, resulting in a runtime of 5 minutes 7 seconds, for a cost of $0.229.

B. `olmOCR-Mix` and `olmOCR-7B` Prompts

B.1. `olmOCR-Mix` construction prompt for GPT-4o

The prompt below was used to create the silver dataset, which we refer to as `olmOCR-Mix` throughout the paper. This dataset consists of structured outputs generated by GPT-4o, using images of PDF pages along with additional layout-aware textual features produced by our DOCUMENT-ANCHORING pipeline. We use this synthetic data to fine-tune our model.

In this prompt, the placeholder `{base_text}` is replaced

[6] `https://www.runpod.io`

with the structured layout-aware text extracted from the PDF using DOCUMENT-ANCHORING. The prompt instructs GPT-4o to output the natural reading-order text of the page, while respecting document semantics, suppressing hallucinations, and formatting content like equations and tables appropriately.

```
Below is the image of one page of a PDF
    document, as well as some raw textual
    content that was previously extracted
    for it that includes position
    information for each image and block of
     text (The origin [0x0] of the
    coordinates is in the lower left corner
     of the image).
Just return the plain text representation
    of this document as if you were reading
     it naturally.
Turn equations into a LaTeX representation,
     and tables into markdown format.
    Remove the headers and footers, but
    keep references and footnotes.
Read any natural handwriting.
This is likely one page out of several in
    the document, so be sure to preserve
    any sentences that come from the
    previous page, or continue onto the
    next page, exactly as they are.
If there is no text at all that you think
    you should read, you can output null.
Do not hallucinate.
RAW_TEXT_START
{base_text}
RAW_TEXT_END
```

B.2. JSON Schema used to prompt GPT-4o

```
"json_schema": {
            "name": "page_response",
            "schema": {
                "type": "object",
                "properties": {
                    "primary_language": {
                        "type": ["string",
                            "null"],
                        "description": "The
                             primary
                            language of the
                             text using two
                            -letter codes
                            or null if
                            there is no
                            text at all
                            that you think
                            you should read
                            .",
                    },
                    "is_rotation_valid": {
                        "type": "boolean",
                        "description": "Is
                            this page
                            oriented
                            correctly for
                            reading? Answer
                             only
                            considering the
                             textual
                            content, do not
                             factor in the
                            rotation of any
                             charts, tables
                            , drawings, or
                            figures.",
                    },
                    "rotation_correction":
                        {
                        "type": "integer",
                        "description": "
                            Indicates the
                            degree of
                            clockwise
                            rotation needed
                             if the page is
                             not oriented
                            correctly.",
                        "enum": [0, 90,
                            180, 270],
                        "default": 0,
                    },
                    "is_table": {
                        "type": "boolean",
                        "description": "
                            Indicates if
                            the majority of
                             the page
                            content is in
                            tabular format
                            .",
                    },
                    "is_diagram": {
                        "type": "boolean",
                        "description": "
                            Indicates if
                            the majority of
                             the page
                            content is a
                            visual diagram
                            .",
                    },
                    "natural_text": {
                        "type": ["string",
                            "null"],
                        "description": "The
                             natural text
                            content
                            extracted from
                            the page.",
                    },
                },
                "additionalProperties":
                    False,
                "required": [
                    "primary_language",
                    "is_rotation_valid",
                    "rotation_correction",
                    "is_table",
                    "is_diagram",
                    "natural_text",
                ],
            },
            "strict": True,
        },
```

B.3. olmOCR-7B prompt

The prompt below is used to draw responses from our fine-tuned model during inference. As before, the placeholder {base_text} is replaced with the output of the DOCUMENT-ANCHORING pipeline i.e., layout-aware textual features extracted from the PDF page.

```
Below is the image of one page of a
    document, as well as some raw textual
    content that was previously extracted
    for it.
Just return the plain text representation
    of this document as if you were reading
     it naturally.
Do not hallucinate.
RAW_TEXT_START
{base_text}
RAW_TEXT_END
```

B.4. olmOCR-Mix Classification Prompt

The prompt and structured schema below was used to classify a sample of documents from olmOCR-Mix.

```
This is an image of a document page, please
     classify it into one of the following
    categories that best overall summarizes
     its nature: academic, legal, brochure,
     slideshow, table, diagram, or other.
    Also determine the primary language of
    the document and your confidence in the
     classification (0-1).
```

```
class DocumentCategory(str, Enum):
    ACADEMIC = "academic"
    LEGAL = "legal"
    BROCHURE = "brochure"
    SLIDESHOW = "slideshow"
    TABLE = "table"
    DIAGRAM = "diagram"
    OTHER = "other"

class DocumentClassification(BaseModel):
    category: DocumentCategory
    language: str
    confidence: float
```

B.5. olmOCR-Mix PII Prompt

We implemented comprehensive prompting for detecting personally identifiable information (PII) in the documents while cleaning the olmOCR-Mix:

```
You are a document analyzer that identifies
     Personally Identifiable Information
(PII) in documents.
Your task is to analyze the provided
    document image and determine:
1. Whether the document is intended for
    public release or dissemination
   (e.g., research paper, public report,
       etc.)
2. If the document contains any PII

For PII identification, follow these
    specific guidelines:
IDENTIFIERS FOR PII:
The following are considered identifiers
    that can make information PII:
- Names (full names, first names, last
    names, nicknames)
- Email addresses
- Phone numbers

PII THAT MUST CO-OCCUR WITH AN IDENTIFIER:
The following types of information should
    ONLY be marked as PII if they occur
ALONGSIDE an identifier (commonly, a person
    's name):
- Addresses (street address, postal code,
    etc.)
- Biographical Information (date of birth,
    place of birth, gender, sexual
  orientation, race, ethnicity, citizenship
      /immigration status, religion)
- Location Information (geolocations,
    specific coordinates)
- Employment Information (job titles,
    workplace names, employment history)
- Education Information (school names,
    degrees, transcripts)
- Medical Information (health records,
    diagnoses, genetic or neural data)

PII THAT OCCURS EVEN WITHOUT AN IDENTIFIER:
The following should ALWAYS be marked as
    PII even if they do not occur
alongside an identifier:
- Government IDs (Social Security Numbers,
    passport numbers, driver's license
  numbers, tax IDs)
- Financial Information (credit card
    numbers, bank account/routing numbers)
- Biometric Data (fingerprints, retina
    scans, facial recognition data,
  voice signatures)
- Login information (ONLY mark as PII when
    a username, password, and login
  location are present together)

If the document is a form, then only
    consider fields which are filled out
with specific values as potential PII.
If this page does not itself contain PII,
    but references documents
(such as curriculum vitae, personal
    statements) that typically contain PII,
then do not mark it as PII.
```

```
Only consider actual occurrences of the PII
     within the document shown.
```

C. Further details of OLMOCR-BENCH

C.1. Prompting Strategies and Implementation Details

This section provides comprehensive documentation of the prompting techniques and design strategies to make OLMOCR-BENCH. These prompting approaches were critical in generating test cases while utilizing LLMs and ensuring consistency across document categories.

C.1.1. MATHEMATICAL EXPRESSIONS

For generating mathematical expression test cases from old scans, we employed direct prompts focused on precision. This concise prompt architecture proved effective in extracting LaTeX representations minimizing hallucination. The explicit instruction to use standard LaTeX delimiters ($$) ensured consistent formatting across the OLMOCR-BENCH.

```
Please extract the mathematical equations
    from the document without
omission. Always output the mathematical
    equations as Latex escaped
with $$. Do not hallucinate.
```

C.1.2. MULTI-COLUMN

For Multi-column documents, we utilized a two-stage prompting strategy. The initial analytical stage established structural context:

```
Analyze this document and provide a
    detailed assessment of its structure.
Focus on the layout, headings, footers, and
     any complex formatting.
Please be precise.
```

This preliminary analysis was incorporated into a subsequent HTML rendering prompt:

```
Render this document as clean, semantic
    HTML. Here is the analysis of the
document structure:

{analysis_text}

Requirements:
1. Use appropriate HTML tags for headings,
    paragraphs, and lists.
2. Use <header> and <footer> for top and
    bottom content.
3. For images, use a placeholder <div> with
     class 'image'.
4. Render math equations inline using \( \)
     or \[ \].
5. Preserve any multi-column layout using
    CSS flexbox or grid.
6. The viewport is fixed at {png_width //
    2}x{png_height // 2} pixels.

Enclose your HTML in a html code block.
```

This approach significantly helped in layout preservation in complex documents by providing explicit dimensional constraints and structural information.

C.1.3. PII DETECTION AND FILTERING

We use the same PII detection and filtering as for construction olmOCR-Mix; see Appendix §B.5.

C.1.4. CLEANING MATHEMATICAL EXPRESSIONS

Mathematical expression verification employed specialized prompting for validating equation presence and accuracy:

```
This is a mathematical expression
    verification task.
I'm showing you a page from a PDF document
    containing mathematical expressions.
Please verify if the following LaTeX
    expression:
{latex_expression}
appears correctly in the document.
Respond with a JSON object containing:
1. "status": "correct" or "incorrect"
2. "confidence": a value between 0 and 1
    representing your confidence in the
    answer
3. "explanation": a brief explanation of
    why you believe the expression is
    correct or incorrect
Focus specifically on checking if this
    exact mathematical expression appears
    in the document.
```

C.1.5. CLEANING READING ORDER TESTS

For natural reading order test cases, we implemented below verification prompt to ensure appropriate text segment relationships:

```
Does the text in the 'before' field and the
     'after' field appear in the same
    region of the page?
Look at the PDF image and determine if
    these texts are located near each other
     or in completely
different parts of the page. Different
    regions could be the captions for
    different images, or
inside of different insets or tables.
    However, appearing the same column of
    text, or in the
naturally flowing next column of text is
    close enough.

Before: {before_text}
```

```
After: {after_text}

Respond with 'YES' if they appear in the
    same region or column, and 'NO' if they
     appear in
different regions. Then explain your
    reasoning in 1-2 sentences.
```

C.1.6. HEADER AND FOOTER VERIFICATION

For validating header and footer text identification, we employed JSON-structured verification prompts:

```
This is a header and footer verification
    task.
I'm showing you a page from a PDF document
    containing headers and footers text.
Please verify if the headers or footers is
    exactly matches the below text.
{header_footer_text}
Respond with a JSON object containing:
1. "status": "correct" or "incorrect"
2. "confidence": a value between 0 and 1
    representing your confidence in the
    answer
3. "explanation": a brief explanation of
    why you believe the text is correct or
    incorrect
Focus specifically on checking if this
    exact header or footer expression
    appears in the document.
```

Our prompting strategy deliberately requested different output formats for different content types (Markdown for general text, LaTeX for equations, HTML for tables) to optimize representation fidelity across diverse document elements. Low temperature settings (typically 0.1) was maintained across all the prompt executions to ensure reproducible outputs, particularly important for establishing consistent test cases.

D. OLMOCR-BENCH Sample Test Classes

Below are are few examples taken from OLMOCR-BENCH

E. Example OLMOCR output

Below are some sample outputs on particularly challenging data. OLMOCR, MinerU, GOT-OCR 2.0 and Marker run with default settings.

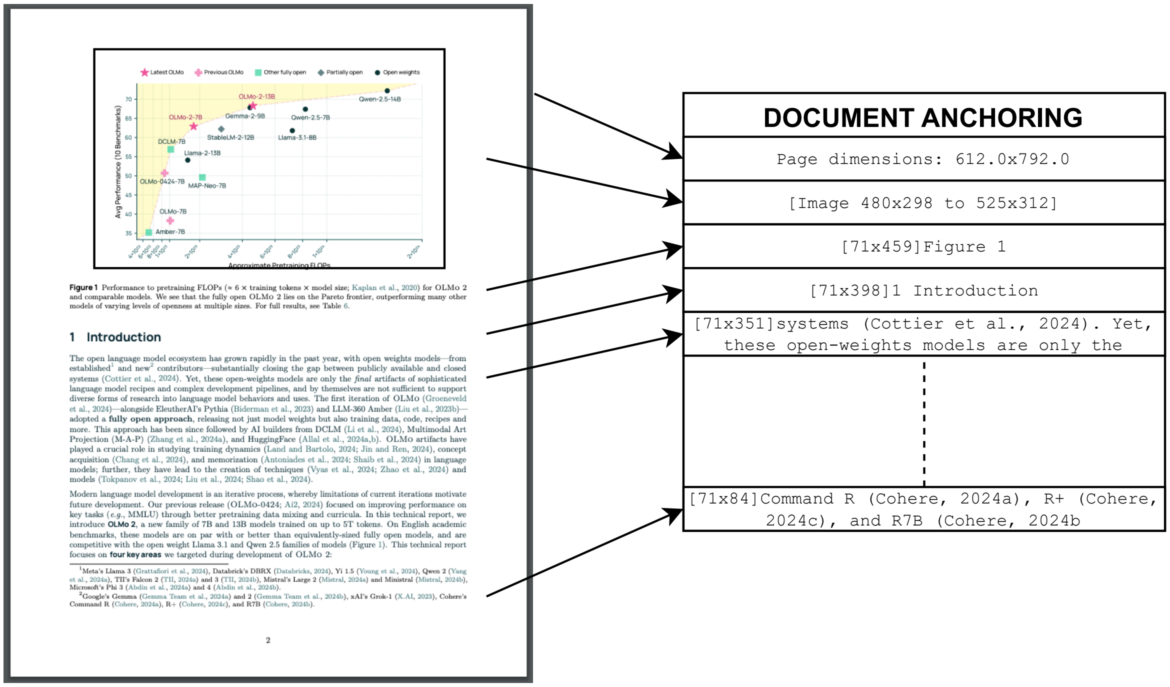


Figure 1 Performance to pretraining FLOPs (≈ 6 × training tokens × model size; Kaplan et al., 2020) for OLMo 2 and comparable models. We see that the fully open OLMo 2 lies on the Pareto frontier, outperforming many other models of varying levels of openness at multiple sizes. For full results, see Table 6.

1 Introduction

The open language model ecosystem has grown rapidly in the past year, with open weights models—from established[1] and new[2] contributors—substantially closing the gap between publicly available and closed systems (Cottier et al., 2024). Yet, these open-weights models are only the *final* artifacts of sophisticated language model recipes and complex development pipelines, and by themselves are not sufficient to support diverse forms of research into language model behaviors and uses. The first iteration of OLMo (Groeneveld et al., 2024)—alongside EleutherAI's Pythia (Biderman et al., 2023) and LLM-360 Amber (Liu et al., 2023b)—adopted a **fully open approach**, releasing not just model weights but also training data, code, recipes and more. This approach has been since followed by AI builders from DCLM (Li et al., 2024), Multimodal Art Projection (M-A-P) (Zhang et al., 2024a), and HuggingFace (Allal et al., 2024a,b). OLMo artifacts have played a crucial role in studying training dynamics (Land and Bartolo, 2024; Jin and Ren, 2024), concept acquisition (Chang et al., 2024), and memorization (Antoniades et al., 2024; Shaib et al., 2024) in language models; further, they have lead to the creation of techniques (Vyas et al., 2024; Zhao et al., 2024) and models (Tokpanov et al., 2024; Liu et al., 2024; Shao et al., 2024).

Modern language model development is an iterative process, whereby limitations of current iterations motivate future development. Our previous release (OLMo-0424; Ai2, 2024) focused on improving performance on key tasks (*e.g.*, MMLU) through better pretraining data mixing and curricula. In this technical report, we introduce **OLMo 2**, a new family of 7B and 13B models trained on up to 5T tokens. On English academic benchmarks, these models are on par with or better than equivalently-sized fully open models, and are competitive with the open weight Llama 3.1 and Qwen 2.5 families of models (Figure 1). This technical report focuses on **four key areas** we targeted during development of OLMo 2:

[1] Meta's Llama 3 (Grattafiori et al., 2024), Databrick's DBRX (Databricks, 2024), Yi 1.5 (Young et al., 2024), Qwen 2 (Yang et al., 2024a), TII's Falcon 2 (TII, 2024a) and 3 (TII, 2024b), Mistral's Large 2 (Mistral, 2024a) and Ministral (Mistral, 2024b), Microsoft's Phi 3 (Abdin et al., 2024a) and 4 (Abdin et al., 2024b).

[2] Google's Gemma (Gemma Team et al., 2024a) and 2 (Gemma Team et al., 2024b), xAI's Grok-1 (X.AI, 2023), Cohere's Command R (Cohere, 2024a), R+ (Cohere, 2024c), and R7B (Cohere, 2024b).

2

```
Page dimensions: 612.0x792.0
[Image 480x298 to 525x312]
[71x459]Figure 1
[71x398]1 Introduction
[71x351]systems (Cottier et al., 2024). Yet,
these open-weights models are only the
[71x84]Command R (Cohere, 2024a), R+ (Cohere,
2024c), and R7B (Cohere, 2024b
```

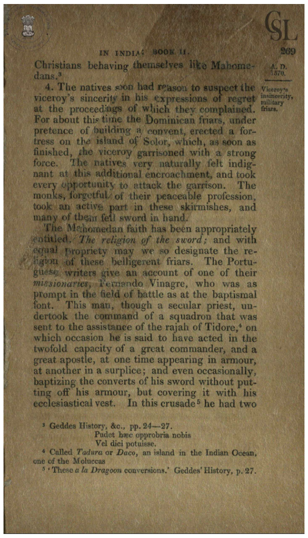
Figure 2. Example of how DOCUMENT-ANCHORING works for a typical page. Relevant image locations and text blocks get extracted, concatenated, and inserted into the model prompt. When prompting a VLM for a plain text version of the document, the anchored text is used in conjunction with the rasterized image of a page.

Table 2. Inference cost comparison against other OCR methods. NVIDIA L40S estimated at $0.79 per hour, H100 80GB estimated at $2.69 per hour. We measured a 12% retry rate for OLMOCR. Full cost breakdown in Appendix A.

Model	Hardware	Tokens/sec	Pages/USD	Cost per million pages
GPT-4o	API	-	80	$12,480
	Batch	-	160	$6,240
Mistral OCR	API	-	1,000	$1,000
MinerU	L40S	238	1,678	$596
Gemini Flash 2	API	-	2,004	$499
	Batch	-	4,008	$249
Marker v1.6.2	L40S	690	4,244	$235
OLMOCR	L40S	906	**5,697**	**$176**
	H100	3,050	**5,632**	**$178**

Table 3. Counts of PDF document types and unit test types in OLMOCR-BENCH.

	Presence	Absence	Read Order	Table	Formula	Total Tests
arXiv Math (AM)	-	-	-	-	2,927	2,927
Old Scans Math (OSM)	-	-	-	-	458	458
Tables (TA)	-	-	-	1,020	-	1,020
Old Scans (OS)	279	70	177	-	-	526
Headers Footers (HF)	-	753	-	-	-	753
Multi Column (MC)	-	-	884	-	-	884
Long Tiny Text (LTT)	442	-	-	-	-	442
Total PDFs	721	823	1,061	1,020	3,385	7,010



IN INDIA: BOOK II. 269

A. D. 1570.

Christians behaving themselves like Mahomedans.[3]

Viceroy's insincerity, military friars.

4. The natives soon had reason to suspect the viceroy's sincerity in his expressions of regret at the proceedings of which they complained. For about this time the Dominican friars, under pretence of building a convent, erected a fortress on the island of Solor, which, as soon as finished, the viceroy garrisoned with a strong force. The natives very naturally felt indignant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of them fell sword in hand.

The Mahomedan faith has been appropriately entitled, *The religion of the sword;* and with equal propriety may we so designate the religion of these belligerent friars. The Portuguese writers give an account of one of their *missionaries,* Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was sent to the assistance of the rajah of Tidore,[4] on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, at another in a surplice; and even occasionally, baptizing the converts of his sword without putting off his armour, but covering it with his ecclesiastical vest. In this crusade[5] he had two

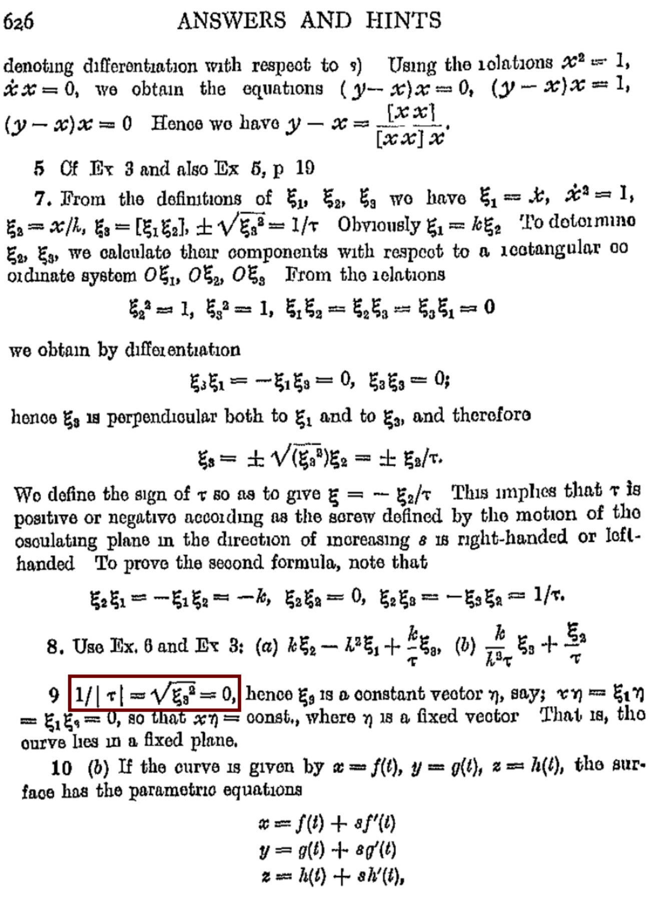
[3] Geddes History, &c., pp. 24—27.

Pudet hæc opprobria nobis
Vel dici potuisse.

[4] Called *Tadura* or *Daco,* an island in the Indian Ocean, one of the Moluccas

[5] 'These *a la Dragoon* conversions.' Geddes' History, p. 27.

OLMOCR	MinerU	GOT-OCR 2.0	Marker
Christians behaving themselves like Mahomedans. 4. The natives soon had reason to suspect the viceroy's sincerity in his expressions of regret at the proceedings of which they complained. For about this time the Dominican friars, under pretence of building a convent, erected a fortress on the island of Solor, which, as soon as finished, the viceroy garrisoned with a strong force. The natives very naturally felt indignant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of them fell sword in hand. The Mahomedan faith has been appropriately entitled, The religion of the sword; and with equal propriety may we so designate the religion of these belligerent friars. The Portuguese writers give an account of one of their missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was sent to the assistance of the rajah of Tidore, on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, at another in a surplice; and even occasionally, baptizing the converts of his sword without putting off his armour, but covering it with his ecclesiastical vest. In this crusade he had two	ININDIASY BOOKU Christians bchaving.themselves like Mahome dans.3 4.The natives soon had reason to suspect ihe viceroy's sincerity in his expressions of regret at the proceedings of which they complained. For about this time the Dominican friars,under pretenceof building a convent,erected a for tress on the island of Solorwhich,as soon as finishedthe viceroy garrisoned with a strong force. The natives very naturally felt indig nant at this additional encroachment, and took every pportunity to attack the garrison.The monks,forgetful of their peaceable profession took an activa part in these skirmishes, and many of tbein feil sword in hand. TheMahornedan faithhas been appropriately ntitled.The religion of the swordand with equal propriety may we so designate the region of these belligerent friars.The Portugueswriters give an account of one of their missionarzes,femando Vinagre,who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was sent to the assistance of the rajah of Tidore,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, at another in a surplice;and even occasionally baptizing the converts of his sword without put ting off his armour, but covering it with his ecclesiastical vest.In this crusadehe had two	IN INDIA: BOOK U 269 Christians behaving themselves like Mahome-1670. 4. The natives son had reason to suspect the Viceroy' s vice roy' s sincerity in his expressions of regret in s in e eri ty at the proceedings of which they complained. fl it ars. For about this time the Dominican f mars, under pre ten ce of building a convent, erected a for- tress on the island of Sol or, which, as soon as finished, the vice roy garrisoned with a strong force. The natives very naturally felt indig- nant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of the n fell sword in hand. The Mh on med an faith has been appropriately entitled. The religion of the sword; and with e ral Tropriety may we so designate the re- gian of these belligerent friars. The Port u- gue s writers give an account of one of their mission are s, Fer endo Vina gre, who was as prompt in the fe ld of battle as at the baptismal font. This man, though a secular priest, un- der took the command of a squadron that was sent to the assistance of the rajah of Tidore, on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, at another in a surplice; and even occasionally, baptizing the converts of his sword without put- ting off his armour, but covering it with his ecclesiastical vest. In this crusade he had two 3 Ged des History, & c. , pp. 24-27. P ude th aec opp rob ria nobis Vel die ipo tui sse. Called Tadur u or Daco, an island in the Indian Ocean, one of the Mol ucc as These a laDra goon conversions. Ged des History, p. 27.	## **IN INDIA *** BOOK TI. S69 Christians behaving themselves like Ma borne- a. dans.3 .“5/0- *t>.* The natives soon had reason to suspect the viceroy, viceroy's sincerity in his expressions of regret at the proceedings of which they complained. “n.“' For about this time the Dominican friars, under pretence of building a. convent, erected a fortress on the island of Sol or, which, as soon as finished, the viceroy garrisoned with a strong force. The natives' very naturally felt indig-S nant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful/ of their peaceable profession, took an active part in these skirmishes, and many of tbg.tr fell sword in hand. The i'lfinomedan faith has been appropriately entitled., 'The religion of the sword',; and with equal propriety may we so designate the re- . i'gv.m of these belligerent friars. The Portugu writers give an account of one of their 'missionaries,' Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was I sent to the assistance of the rajah of Tidore,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, ; at another in a surplice; and even occasionally, baptizing the converts of his sword without putting off his armour, but covering it with his ecclesiastical vest. In this crusade5 he had two 3 Geddes History, &c., pp. 24—27. Pudet hæc opprobria nobis Vel dici potuisse. 4 Called 'T a d u ra' or 'D a c o,' an island in the Indian Ocean, one of the Moluccas 5 'These 'a la D ra g o o n' conversions.' Geddes' History, p. 27.



626 ANSWERS AND HINTS

denoting differentiation with respect to s) Using the relations $x^2 = 1$, $\dot{x}x = 0$, we obtain the equations $(y - x)x = 0$, $(y - x)x = 1$, $(y - x)x = 0$ Hence we have $y - x = \frac{[x x]}{[x x] x}$.

5 Cf Ex 3 and also Ex 5, p 19

7. From the definitions of ξ_1, ξ_2, ξ_3 we have $\xi_1 = \dot{x}$, $\dot{x}^2 = 1$, $\xi_2 = x/k$, $\xi_3 = [\xi_1 \xi_2]$, $\pm\sqrt{\xi_3^2} = 1/\tau$ Obviously $\xi_1 = k\xi_2$ To determine ξ_2, ξ_3, we calculate their components with respect to a rectangular coordinate system $O\xi_1$, $O\xi_2$, $O\xi_3$ From the relations

$$\xi_2^2 = 1,\ \xi_3^2 = 1,\ \xi_1\xi_2 = \xi_2\xi_3 = \xi_3\xi_1 = 0$$

we obtain by differentiation

$$\xi_3\xi_1 = -\xi_1\xi_3 = 0,\ \xi_3\xi_3 = 0;$$

hence ξ_3 is perpendicular both to ξ_1 and to ξ_3, and therefore

$$\xi_3 = \pm\sqrt{(\xi_3^2)}\xi_2 = \pm\,\xi_2/\tau.$$

We define the sign of τ so as to give $\xi = -\,\xi_2/\tau$ This implies that τ is positive or negative according as the screw defined by the motion of the osculating plane in the direction of increasing s is right-handed or left-handed To prove the second formula, note that

$$\xi_2\xi_1 = -\xi_1\xi_2 = -k,\ \xi_2\xi_2 = 0,\ \xi_2\xi_3 = -\xi_3\xi_2 = 1/\tau.$$

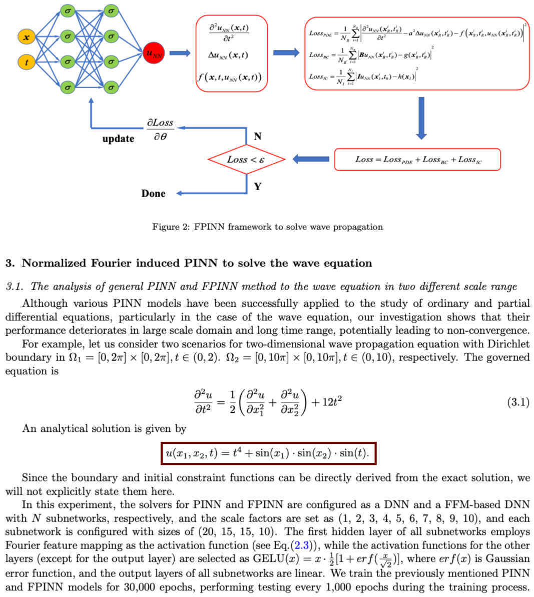
8. Use Ex. 6 and Ex 3: (a) $k\xi_2 - k^2\xi_1 + \frac{k}{\tau}\xi_3$, (b) $\frac{k}{k^3\tau}\,\xi_3 + \frac{\xi_2}{\tau}$

9 $\boxed{1/|\,\tau| = \sqrt{\xi_3^2} = 0,}$ hence ξ_3 is a constant vector η, say; $x\eta = \xi_1\eta = \xi_1\xi_3 = 0$, so that $x\eta =$ const., where η is a fixed vector That is, the curve lies in a fixed plane.

10 (b) If the curve is given by $x = f(t)$, $y = g(t)$, $z = h(t)$, the surface has the parametric equations

$$x = f(t) + sf'(t)$$
$$y = g(t) + sg'(t)$$
$$z = h(t) + sh'(t),$$

Figure 3. Sample visualization from old_scans_math. The OCR output for the highlighted equation should be: `1/|\tau| = \sqrt{\xi_{3}^{2}} = 0`



x
t
σ
u_{NN}
$\frac{\partial^2 u_{NN}(x,t)}{\partial t^2}$
$\Delta u_{NN}(x,t)$
$f(x,t,u_{NN}(x,t))$
$Loss_{PDE} = \frac{1}{N_R}\sum_{i=1}^{N_R}\left|\frac{\partial^2 u_{NN}(x_R^i,t_R^i)}{\partial t^2} - a^2\Delta u_{NN}(x_R^i,t_R^i) - f(x_R^i,t_R^i,u_{NN}(x_R^i,t_R^i))\right|^2$
$Loss_{BC} = \frac{1}{N_B}\sum_{i=1}^{N_B}\left|Bu_{NN}(x_B^i,t_B^i) - g(x_B^i,t_B^i)\right|^2$
$Loss_{IC} = \frac{1}{N_I}\sum_{i=1}^{N_I}\left|Iu_{NN}(x_I^i,t_0) - h(x_I)\right|^2$
$Loss = Loss_{PDE} + Loss_{BC} + Loss_{IC}$
$Loss < \varepsilon$
N
Y
update
$\frac{\partial Loss}{\partial \theta}$
Done

Figure 2: FPINN framework to solve wave propagation

3. Normalized Fourier induced PINN to solve the wave equation

3.1. The analysis of general PINN and FPINN method to the wave equation in two different scale range

Although various PINN models have been successfully applied to the study of ordinary and partial differential equations, particularly in the case of the wave equation, our investigation shows that their performance deteriorates in large scale domain and long time range, potentially leading to non-convergence.

For example, let us consider two scenarios for two-dimensional wave propagation equation with Dirichlet boundary in $\Omega_1 = [0, 2\pi] \times [0, 2\pi], t \in (0, 2)$. $\Omega_2 = [0, 10\pi] \times [0, 10\pi], t \in (0, 10)$, respectively. The governed equation is

$$\frac{\partial^2 u}{\partial t^2} = \frac{1}{2}\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) + 12t^2 \tag{3.1}$$

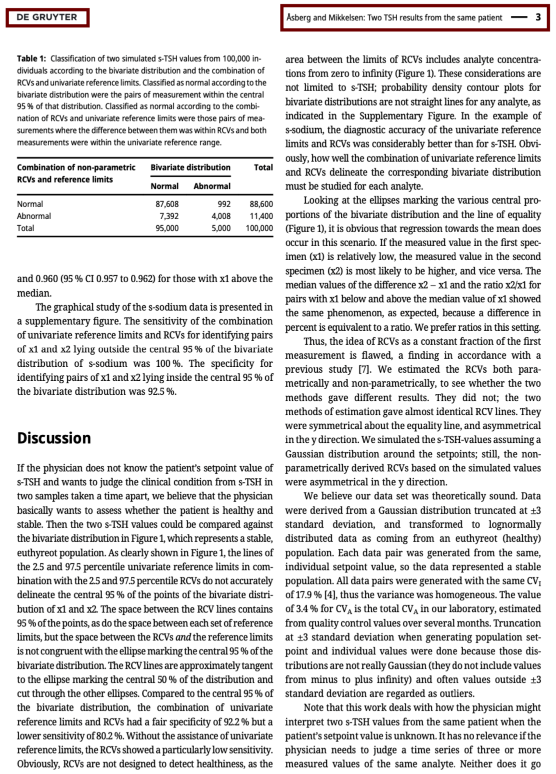
An analytical solution is given by

$$\boxed{u(x_1, x_2, t) = t^4 + \sin(x_1) \cdot \sin(x_2) \cdot \sin(t).}$$

Since the boundary and initial constraint functions can be directly derived from the exact solution, we will not explicitly state them here.

In this experiment, the solvers for PINN and FPINN are configured as a DNN and a FFM-based DNN with N subnetworks, respectively, and the scale factors are set as (1, 2, 3, 4, 5, 6, 7, 8, 9, 10), and each subnetwork is configured with sizes of (20, 15, 15, 10). The first hidden layer of all subnetworks employs Fourier feature mapping as the activation function (see Eq.(2.3)), while the activation functions for the other layers (except for the output layer) are selected as $\mathrm{GELU}(x) = x \cdot \frac{1}{2}[1 + erf(\frac{x}{\sqrt{2}})]$, where $erf(x)$ is Gaussian error function, and the output layers of all subnetworks are linear. We train the previously mentioned PINN and FPINN models for 30,000 epochs, performing testing every 1,000 epochs during the training process.

Figure 4. Sample visualization of a math equation from arXiv_math. The OCR output for the highlighted equation should be: `u(x_{1},x_{2},t)=t^{4} + \text{sin}(x_{1}) \cdot \text{sin}(x_{2}) \cdot \text{sin}(t)`



DE GRUYTER

Åsberg and Mikkelsen: Two TSH results from the same patient — 3

Table 1: Classification of two simulated s-TSH values from 100,000 individuals according to the bivariate distribution and the combination of RCVs and univariate reference limits. Classified as normal according to the bivariate distribution were the pairs of measurement within the central 95 % of that distribution. Classified as normal according to the combination of RCVs and univariate reference limits were those pairs of measurements where the difference between them was within RCVs and both measurements were within the univariate reference range.

Combination of non-parametric RCVs and reference limits	Bivariate distribution		Total
	Normal	Abnormal	
Normal	87,608	992	88,600
Abnormal	7,392	4,008	11,400
Total	95,000	5,000	100,000

and 0.960 (95 % CI 0.957 to 0.962) for those with x1 above the median.

The graphical study of the s-sodium data is presented in a supplementary figure. The sensitivity of the combination of univariate reference limits and RCVs for identifying pairs of x1 and x2 lying outside the central 95 % of the bivariate distribution of s-sodium was 100 %. The specificity for identifying pairs of x1 and x2 lying inside the central 95 % of the bivariate distribution was 92.5 %.

Discussion

If the physician does not know the patient's setpoint value of s-TSH and wants to judge the clinical condition from s-TSH in two samples taken a time apart, we believe that the physician basically wants to assess whether the patient is healthy and stable. Then the two s-TSH values could be compared against the bivariate distribution in Figure 1, which represents a stable, euthyreot population. As clearly shown in Figure 1, the lines of the 2.5 and 97.5 percentile univariate reference limits in combination with the 2.5 and 97.5 percentile RCVs do not accurately delineate the central 95 % of the points of the bivariate distribution of x1 and x2. The space between the RCV lines contains 95 % of the points, as do the space between each set of reference limits, but the space between the RCVs *and* the reference limits is not congruent with the ellipse marking the central 95 % of the bivariate distribution. The RCV lines are approximately tangent to the ellipse marking the central 50 % of the distribution and cut through the other ellipses. Compared to the central 95 % of the bivariate distribution, the combination of univariate reference limits and RCVs had a fair specificity of 92.2 % but a lower sensitivity of 80.2 %. Without the assistance of univariate reference limits, the RCVs showed a particularly low sensitivity. Obviously, RCVs are not designed to detect healthiness, as the area between the limits of RCVs includes analyte concentrations from zero to infinity (Figure 1). These considerations are not limited to s-TSH; probability density contour plots for bivariate distributions are not straight lines for any analyte, as indicated in the Supplementary Figure. In the example of s-sodium, the diagnostic accuracy of the univariate reference limits and RCVs was considerably better than for s-TSH. Obviously, how well the combination of univariate reference limits and RCVs delineate the corresponding bivariate distribution must be studied for each analyte.

Looking at the ellipses marking the various central proportions of the bivariate distribution and the line of equality (Figure 1), it is obvious that regression towards the mean does occur in this scenario. If the measured value in the first specimen (x1) is relatively low, the measured value in the second specimen (x2) is most likely to be higher, and vice versa. The median values of the difference x2 – x1 and the ratio x2/x1 for pairs with x1 below and above the median value of x1 showed the same phenomenon, as expected, because a difference in percent is equivalent to a ratio. We prefer ratios in this setting.

Thus, the idea of RCVs as a constant fraction of the first measurement is flawed, a finding in accordance with a previous study [7]. We estimated the RCVs both parametrically and non-parametrically, to see whether the two methods gave different results. They did not; the two methods of estimation gave almost identical RCV lines. They were symmetrical about the equality line, and asymmetrical in the y direction. We simulated the s-TSH-values assuming a Gaussian distribution around the setpoints; still, the non-parametrically derived RCVs based on the simulated values were asymmetrical in the y direction.

We believe our data set was theoretically sound. Data were derived from a Gaussian distribution truncated at ±3 standard deviation, and transformed to lognormally distributed data as coming from an euthyreot (healthy) population. Each data pair was generated from the same, individual setpoint value, so the data represented a stable population. All data pairs were generated with the same CV_I of 17.9 % [4], thus the variance was homogeneous. The value of 3.4 % for CV_A is the total CV_A in our laboratory, estimated from quality control values over several months. Truncation at ±3 standard deviation when generating population setpoint and individual values were done because those distributions are not really Gaussian (they do not include values from minus to plus infinity) and often values outside ±3 standard deviation are regarded as outliers.

Note that this work deals with how the physician might interpret two s-TSH values from the same patient when the patient's setpoint value is unknown. It has no relevance if the physician needs to judge a time series of three or more measured values of the same analyte. Neither does it go

Figure 5. Sample visualization of `headers_footers`. We want the OCR to skip the document headers and page number.

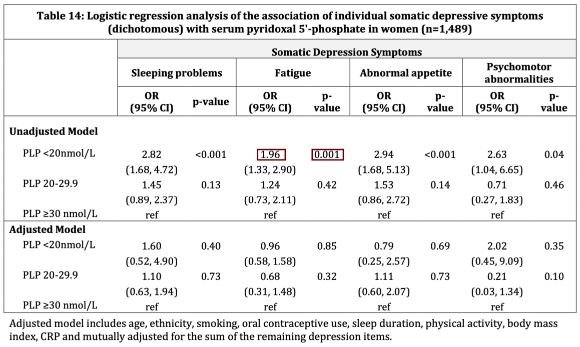
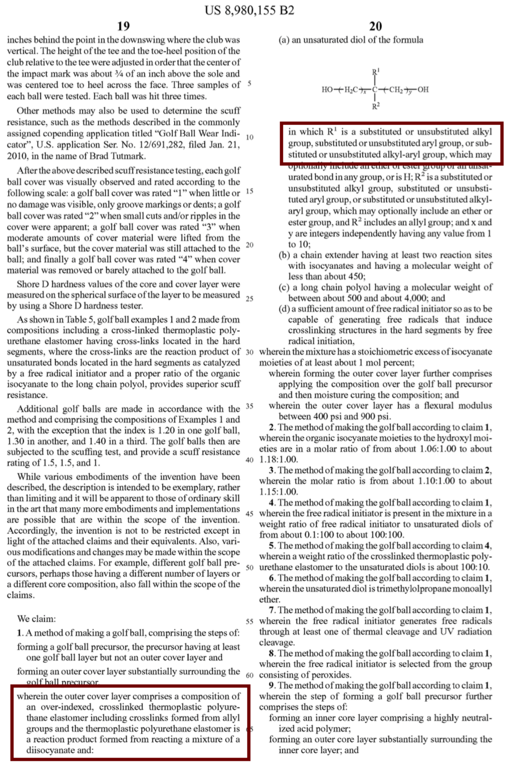


Table 14: Logistic regression analysis of the association of individual somatic depressive symptoms (dichotomous) with serum pyridoxal 5'-phosphate in women (n=1,489)

	Somatic Depression Symptoms							
	Sleeping problems		Fatigue		Abnormal appetite		Psychomotor abnormalities	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Unadjusted Model								
PLP <20nmol/L	2.82 (1.68, 4.72)	<0.001	1.96 (1.33, 2.90)	0.001	2.94 (1.68, 5.13)	<0.001	2.63 (1.04, 6.65)	0.04
PLP 20-29.9	1.45 (0.89, 2.37)	0.13	1.24 (0.73, 2.11)	0.42	1.53 (0.86, 2.72)	0.14	0.71 (0.27, 1.83)	0.46
PLP ≥30 nmol/L	ref		ref		ref		ref	
Adjusted Model								
PLP <20nmol/L	1.60 (0.52, 4.90)	0.40	0.96 (0.58, 1.58)	0.85	0.79 (0.25, 2.57)	0.69	2.02 (0.45, 9.09)	0.35
PLP 20-29.9	1.10 (0.63, 1.94)	0.73	0.68 (0.31, 1.48)	0.32	1.11 (0.60, 2.07)	0.73	0.21 (0.03, 1.34)	0.10
PLP ≥30 nmol/L	ref		ref		ref		ref	

Adjusted model includes age, ethnicity, smoking, oral contraceptive use, sleep duration, physical activity, body mass index, CRP and mutually adjusted for the sum of the remaining depression items.

Figure 6. Sample visualization of `table_tests`. We want the OCR to predict that cell 1.96 is to the left of cell 0.001.



US 8,980,155 B2

19

inches behind the point in the downswing where the club was vertical. The height of the tee and the toe-heel position of the club relative to the tee were adjusted in order that the center of the impact mark was about ¾ of an inch above the sole and was centered toe to heel across the face. Three samples of each ball were tested. Each ball was hit three times.

Other methods may also be used to determine the scuff resistance, such as the methods described in the commonly assigned copending application titled "Golf Ball Wear Indicator", U.S. application Ser. No. 12/691,282, filed Jan. 21, 2010, in the name of Brad Tutmark.

After the above described scuff resistance testing, each golf ball cover was visually observed and rated according to the following scale: a golf ball cover was rated "1" when little or no damage was visible, only groove markings or dents; a golf ball cover was rated "2" when small cuts and/or ripples in the cover were apparent; a golf ball cover was rated "3" when moderate amounts of cover material were lifted from the ball's surface, but the cover material was still attached to the ball; and finally a golf ball cover was rated "4" when cover material was removed or barely attached to the golf ball.

Shore D hardness values of the core and cover layer were measured on the spherical surface of the layer to be measured by using a Shore D hardness tester.

As shown in Table 5, golf ball examples 1 and 2 made from compositions including a cross-linked thermoplastic polyurethane elastomer having cross-links located in the hard segments, where the cross-links are the reaction product of unsaturated bonds located in the hard segments as catalyzed by a free radical initiator and a proper ratio of the organic isocyanate to the long chain polyol, provides superior scuff resistance.

Additional golf balls are made in accordance with the method and comprising the compositions of Examples 1 and 2, with the exception that the index is 1.20 in one golf ball, 1.30 in another, and 1.40 in a third. The golf balls then are subjected to the scuffing test, and provide a scuff resistance rating of 1.5, 1.5, and 1.

While various embodiments of the invention have been described, the description is intended to be exemplary, rather than limiting and it will be apparent to those of ordinary skill in the art that many more embodiments and implementations are possible that are within the scope of the invention. Accordingly, the invention is not to be restricted except in light of the attached claims and their equivalents. Also, various modifications and changes may be made within the scope of the attached claims. For example, different golf ball precursors, perhaps those having a different number of layers or a different core composition, also fall within the scope of the claims.

We claim:

1. A method of making a golf ball, comprising the steps of:

forming a golf ball precursor, the precursor having at least one golf ball layer but not an outer cover layer and

forming an outer cover layer substantially surrounding the golf ball precursor,

wherein the outer cover layer comprises a composition of an over-indexed, crosslinked thermoplastic polyurethane elastomer including crosslinks formed from allyl groups and the thermoplastic polyurethane elastomer is a reaction product formed from reacting a mixture of a diisocyanate and:

20

(a) an unsaturated diol of the formula

in which R^1 is a substituted or unsubstituted alkyl group, substituted or unsubstituted aryl group, or substituted or unsubstituted alkyl-aryl group, which may optionally include an ether or ester group or an unsaturated bond in any group, or is H; R^2 is a substituted or unsubstituted alkyl group, substituted or unsubstituted aryl group, or substituted or unsubstituted alkyl-aryl group, which may optionally include an ether or ester group, and R^2 includes an allyl group; and x and y are integers independently having any value from 1 to 10;

(b) a chain extender having at least two reaction sites with isocyanates and having a molecular weight of less than about 450;

(c) a long chain polyol having a molecular weight of between about 500 and about 4,000; and

(d) a sufficient amount of free radical initiator so as to be capable of generating free radicals that induce crosslinking structures in the hard segments by free radical initiation,

wherein the mixture has a stoichiometric excess of isocyanate moieties of at least about 1 mol percent;

wherein forming the outer cover layer further comprises applying the composition over the golf ball precursor and then moisture curing the composition; and

wherein the outer cover layer has a flexural modulus between 400 psi and 900 psi.

2. The method of making the golf ball according to claim **1**, wherein the organic isocyanate moieties to the hydroxyl moieties are in a molar ratio of from about 1.06:1.00 to about 1.18:1.00.

3. The method of making the golf ball according to claim **2**, wherein the molar ratio is from about 1.10:1.00 to about 1.15:1.00.

4. The method of making the golf ball according to claim **1**, wherein the free radical initiator is present in the mixture in a weight ratio of free radical initiator to unsaturated diols of from about 0.1:100 to about 100:100.

5. The method of making the golf ball according to claim **4**, wherein a weight ratio of the crosslinked thermoplastic polyurethane elastomer to the unsaturated diols is about 100:10.

6. The method of making the golf ball according to claim **1**, wherein the unsaturated diol is trimethylolpropane monoallyl ether.

7. The method of making the golf ball according to claim **1**, wherein the free radical initiator generates free radicals through at least one of thermal cleavage and UV radiation cleavage.

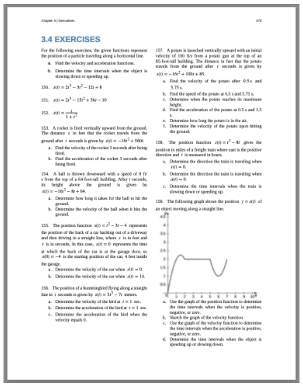
8. The method of making the golf ball according to claim **1**, wherein the free radical initiator is selected from the group consisting of peroxides.

9. The method of making the golf ball according to claim **1**, wherein the step of forming a golf ball precursor further comprises the steps of:

forming an inner core layer comprising a highly neutralized acid polymer;

forming an outer core layer substantially surrounding the inner core layer; and

Figure 7. Sample visualization of `reading_order`. The reading order should start with the left column before moving to the right column.



Chapter 3 | Derivatives 273

3.4 EXERCISES

For the following exercises, the given functions represent the position of a particle traveling along a horizontal line.

a. Find the velocity and acceleration functions.

b. Determine the time intervals when the object is slowing down or speeding up.

150. $s(t) = 2t^3 - 3t^2 - 12t + 8$

151. $s(t) = 2t^3 - 15t^2 + 36t - 10$

152. $s(t) = \frac{t}{1+t^2}$

153. A rocket is fired vertically upward from the ground. The distance s in feet that the rocket travels from the ground after t seconds is given by $s(t) = -16t^2 + 560t$.

a. Find the velocity of the rocket 3 seconds after being fired.

b. Find the acceleration of the rocket 3 seconds after being fired.

154. A ball is thrown downward with a speed of 8 ft/s from the top of a 64-foot-tall building. After t seconds, its height above the ground is given by $s(t) = -16t^2 - 8t + 64$.

a. Determine how long it takes for the ball to hit the ground.

b. Determine the velocity of the ball when it hits the ground.

155. The position function $s(t) = t^2 - 3t - 4$ represents the position of the back of a car backing out of a driveway and then driving in a straight line, where s is in feet and t is in seconds. In this case, $s(t) = 0$ represents the time at which the back of the car is at the garage door, so $s(0) = -4$ is the starting position of the car, 4 feet inside the garage.

a. Determine the velocity of the car when $s(t) = 0$.

b. Determine the velocity of the car when $s(t) = 14$.

156. The position of a hummingbird flying along a straight line in t seconds is given by $s(t) = 3t^3 - 7t$ meters.

a. Determine the velocity of the bird at $t = 1$ sec.

b. Determine the acceleration of the bird at $t = 1$ sec.

c. Determine the acceleration of the bird when the velocity equals 0.

157. A potato is launched vertically upward with an initial velocity of 100 ft/s from a potato gun at the top of an 85-foot-tall building. The distance in feet that the potato travels from the ground after t seconds is given by $s(t) = -16t^2 + 100t + 85$.

a. Find the velocity of the potato after 0.5 s and 5.75 s.

b. Find the speed of the potato at 0.5 s and 5.75 s.

c. Determine when the potato reaches its maximum height.

d. Find the acceleration of the potato at 0.5 s and 1.5 s.

e. Determine how long the potato is in the air.

f. Determine the velocity of the potato upon hitting the ground.

158. The position function $s(t) = t^3 - 8t$ gives the position in miles of a freight train where east is the positive direction and t is measured in hours.

a. Determine the direction the train is traveling when $s(t) = 0$.

b. Determine the direction the train is traveling when $a(t) = 0$.

c. Determine the time intervals when the train is slowing down or speeding up.

159. The following graph shows the position $y = s(t)$ of an object moving along a straight line.

a. Use the graph of the position function to determine the time intervals when the velocity is positive, negative, or zero.

b. Sketch the graph of the velocity function.

c. Use the graph of the velocity function to determine the time intervals when the acceleration is positive, negative, or zero.

d. Determine the time intervals when the object is speeding up or slowing down.

OLMOCR

3.4 EXERCISES

For the following exercises, the given functions represent the position of a particle traveling along a horizontal line.

a. Find the velocity and acceleration functions.

b. Determine the time intervals when the object is slowing down or speeding up.

150. $s(t) = 2t^3 - 3t^2 - 12t + 8$

151. $s(t) = 2t^3 - 15t^2 + 36t - 10$

152. $s(t) = \frac{t}{1+t^2}$

153. A rocket is fired vertically upward from the ground. The distance s in feet that the rocket travels from the ground after t seconds is given by $s(t) = -16t^2 + 560t$.

a. Find the velocity of the rocket 3 seconds after being fired.

b. Find the acceleration of the rocket 3 seconds after being fired.

154. A ball is thrown downward with a speed of 8 ft/s from the top of a 64-foot-tall building. After t seconds, its height above the ground is given by $s(t) = -16t^2 - 8t + 64$.

a. Determine how long it takes for the ball to hit the ground.

b. Determine the velocity of the ball when it hits the ground.

155. The position function $s(t) = t^2 - 3t - 4$ represents the position of the back of a car backing out of a driveway and then driving in a straight line, where s is in feet and t is in seconds. In this case, $s(t) = 0$ represents the time at which the back of the car is at the garage door, so $s(0) = -4$ is the starting position of the car, 4 feet inside the garage.

a. Determine the velocity of the car when $s(t) = 0$.

b. Determine the velocity of the car when $s(t) = 14$.

156. The position of a hummingbird flying along a straight line in t seconds is given by $s(t) = 3t^3 - 7t$ meters.

a. Determine the velocity of the bird at $t = 1$ sec.

b. Determine the acceleration of the bird at $t = 1$ sec.

c. Determine the acceleration of the bird when the velocity equals 0.

157. A potato is launched vertically upward with an initial velocity of 100 ft/s from a potato gun at the top of an 85-foot-tall building. The distance in feet that the potato travels from the ground after t seconds is given by $s(t) = -16t^2 + 100t + 85$. . . .

MinerU

# 3.4 EXERCISES

For the following exercises, the given functions represent the position of a particle traveling along a horizontal line.

a. Find the velocity and acceleration functions. b. Determine the time intervals when the object is slowing down or speeding up.

150. $s(t) = 2t^3 - 3t^2 - 12t + 8$ 151. $s(t) = 2t^3 - 15t^2 + 36t - 10$ 152. $s(t) = \frac{t}{1+t^2}$

153. A rocket is fired vertically upward from the ground. The distance s in feet that the rocket travels from the ground after t seconds is given by $s(t) = -16t^2 + 560t, .$

a. Find the velocity of the rocket 3 seconds after being fired. b. Find the acceleration of the rocket 3 seconds after being fired.

154. A ball is thrown downward with a speed of 8 ft/ s from the top of a 64-foot-tall building. After t seconds, its height above the ground is given by $s(t) = -16t^2 - 8t + 64.$.

a. Determine how long it takes for the ball to hit the ground. b. Determine the velocity of the ball when it hits the ground.

155. The position function $s(t) = t^2 - 3t - 4$ represents the position of the back of a car backing out of a driveway and then driving in a straight line, where s is in feet and t is in seconds. In this case, $s(t) = 0$ represents the time at which the back of the car is at the garage door, so $s(0) = -4$ is the starting position of the car, 4 feet inside the garage.

a. Determine the velocity of the car when $s(t) = 0$. b. Determine the velocity of the car when $s(t) = 14$.

156. The position of a hummingbird flying along a straight line in t seconds is given by $s(t) = 3t^3 - 7t$ meters.

a. Determine the velocity of the bird at $t = 1$ sec. b. Determine the acceleration of the bird at $t = 1$ sec. c. Determine the acceleration of the bird when the velocity equals 0.

157. A potato is launched vertically upward with an initial velocity of 100 ft/s from a potato gun at the top of an 85-foot-tall building. The distance in feet that the potato travels from the ground after t seconds is given by $s(t) = -16t^2 + 100t + 85.$.

. . .

GOT-OCR 2.0

Chapter 3 | Derivatives 273 3.4 EXERCISES For the following exercises, the given functions represent the position of a particle traveling along a horizontal line. a. Find the velocity and acceleration functions. b. Determine the time intervals when the object is slowing down or speeding up. 150. s(t) = 2t3 -3t2 -12t + 8 151. s(t) = 2t3 -15t2 + 36t -10 152. s(t) = t 1 + t2 153. A rocket is fired vertically upward from the ground. The distance s in feet that the rocket travels from the ground after t seconds is given by s(t) = -16t2 + 560t. a. Find the velocity of the rocket 3 seconds after being fired. b. Find the acceleration of the rocket 3 seconds after being fired. 154. A ball is thrown downward with a speed of 8 ft/ s from the top of a 64-foot-tall building. After t seconds, its height above the ground is given by s(t) = - 16t2 - 8t + 64. a. Determine how long it takes for the ball to hit the ground. b. Determine the velocity of the ball when it hits the ground. 155. The position function s(t) = t2 - 3t - 4 represents the position of the back of a car backing out of a driveway and then driving in a straight line, where s is in feet and t is in seconds. In this case, s(t) = 0 represents the time at which the back of the car is at the garage door, so s(0) = - 4 is the starting position of the car, 4 feet inside the garage. a. Determine the velocity of the car when s(t) = 0. b. Determine the velocity of the car when s(t) = 14. 156. The position of a hummingbird flying along a straight line in t seconds is given by s(t) = 3t3 - 7t 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 2 2 2 2 2 2 2 2 2 0 1 1 1 1 1 1 1 1 3 4 4 4 4 4 4 4 4 3 4 4 4 4 4 4 4 4 2 3 3 3 3 3 3 3 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 0 2 2 2 2 2 2 2 2 5 5 5 5 5 5 5 5 5 1 1 1 1 1 1 1 1 5 5 5 5 5 5 5 5 5 0 0 0 0 0 0 0 0 5 5 5 5 5 5 5 5 5 3 3 3 3 3 3 3 3 3 5 5 5 5 5 5 5 5 5 2 2 2 2 2 2 2 2 2 a. Use the graph of the position function to determine the time intervals when the velocity is positive, negative, or zero. b. Sketch the graph of the velocity function. c. Use the graph of the velocity function to determine the time intervals when the acceleration is positive, negative, or zero. d. Determine the time intervals when the object is speeding up or slowing down. . . .

Marker

## **3.4 EXERCISES**

For the following exercises, the given functions represent the position of a particle traveling along a horizontal line.

- a. Find the velocity and acceleration functions. - b. Determine the time intervals when the object is slowing down or speeding up.

$$150. \quad s(t) = 2t^3 - 3t^2 - 12t + 8$$

$$151. \quad s(t) = 2t^3 - 15t^2 + 36t - 10t$$

$$152. \quad s(t) = \frac{t}{1+t^2}$$

153. A rocket is fired vertically upward from the ground. The distance *s* in feet that the rocket travels from the ground after *t* seconds is given by *s*(*t*) = -16*t* 2 + 560*t*.

- a. Find the velocity of the rocket 3 seconds after being fired. - b. Find the acceleration of the rocket 3 seconds after being fired.

154. A ball is thrown downward with a speed of 8 ft/ s from the top of a 64-foot-tall building. After *t* seconds, its height above the ground is given by *s*(*t*) = -16*t* 2 - 8*t* + 64.

- a. Determine how long it takes for the ball to hit the ground. - b. Determine the velocity of the ball when it hits the ground.

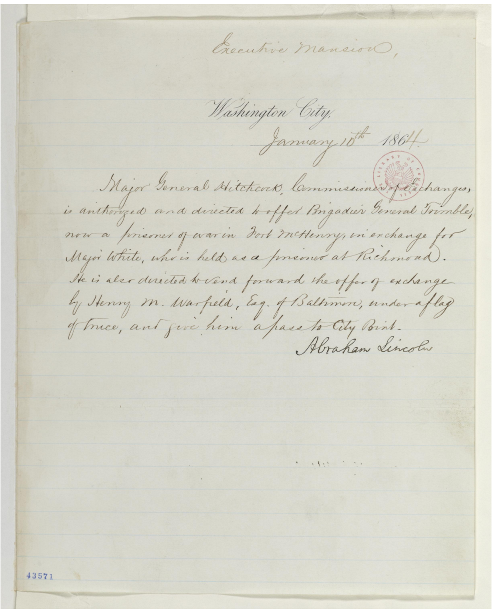
155. The position function *s*(*t*) = *t* 2 - 3*t* - 4 represents the position of the back of a car backing out of a driveway and then driving in a straight line, where *s* is in feet and *t* is in seconds. In this case, *s*(*t*) = 0 represents the time at which the back of the car is at the garage door, so *s*(0) = -4 is the starting position of the car, 4 feet inside the garage.

- a. Determine the velocity of the car when *s*(*t*) = 0. - b. Determine the velocity of the car when *s*(*t*) = 14.

156. The position of a hummingbird flying along a straight line in *t* seconds is given by *s*(*t*) = 3*t* 3 - 7*t* meters.

- a. Determine the velocity of the bird at *t* = 1 sec. - b. Determine the acceleration of the bird at *t* = 1 sec. - c. Determine the acceleration of the bird when the velocity equals 0.

. . .



Executive Mansion,

Washington City,
January 10th 1864

Major General Hitchcock, Commissioner of Exchanges, is authorized and directed to offer Brigadier General Trimble, now a prisoner of war in Fort McHenry, in exchange for Major White, who is held as a prisoner at Richmond. He is also directed to send forward the offer of exchange by Henry M. Warfield, Esq. of Baltimore, under a flag of truce, and give him a pass to City Point.

Abraham Lincoln

olmOCR	MinerU	GOT-OCR 2.0	Marker
Executive Mansion, Washington City, January 15th, 1864 Major General Hitchcock, Commissioner of Exchanges, is authorized and directed to offer Brigadier General Trimble, now a prisoner of war in Fort McHenry, in exchange for Major White, who is held as a prisoner at Richmond. He is also directed to send forward the offer of exchange by Henry M. Warfield, Esq. of Baltimore, under a flag of truce, and give him a pass to City Point. Abraham Lincoln	*No text produced.*	43571 Bachington City January 10th 1864. Major General Architect, Commissioner of aivachangera is authorized and directed by ffeed Bngader General Trelmble, new a firemen of war in Fert nchery in exchange for Mayor White, who held a a firemen at Hillmannd. He is aker conducted by end forward the offer of exchange by Henry in. Warfield, Lag. of Balthmore, under a flag of three, and five him afaies to City Bink. Abraham Lincoln	necuhve Mansion Vastington amany layor Seneral Hitchcocks Commissioner of Cachanges, is anthonged and directed to offer Bingadier General Trin prisoner of war in Fort Inctienny, in exchange now w Major White, who is held as a preises at Richmond Ite is also directed to vand forwards the offer of exchange by Stenny in. Warfield, Eag. of Baltimore, under aflag 11 mice, and give him apass to tity Point. Abrakan Sincolus