Staged Continual Adaptation of Multimodal Foundation Models for Japanese Financial Documents

Genshin Kakimoto [1] **Atsushi Yanagisawa** [1]

Abstract

Staged post-training of multimodal foundation models chains vision–language alignment, reasoning distillation, and domain tuning, yet how each phase trades capabilities against the others remains under-characterised. Tracking an 8.4B-parameter model from an un-fine-tuned baseline through three training phases on Japanese financial documents (evaluated on five text-input benchmarks), we find the phases interact *non-additively*: math-only distillation overwrites the financial knowledge acquired earlier even as procedural math skill survives, and the domain phase repairs more than it installs. A skip-phase ablation is best for domain knowledge but forfeits numerical-reasoning gains, so per-phase contributions are not decomposable and the best checkpoint is task-dependent, with implications for checkpoint retention and scheduling rather than added stages.

1. Introduction

Adapting a multimodal foundation model to a specialised domain rarely proceeds in a single step: contemporary recipes typically combine vision–language alignment, reasoning distillation, and domain tuning, where each stage can both install new capabilities and erode previously acquired ones. LLaVA-style recipes (Liu et al., 2023; Ye et al., 2023; Zhang et al., 2023; Li et al., 2024; Zhou et al., 2025) establish the architectural template for such multi-stage adaptation but say little about how earlier phases interact with later ones when the target domain demands structured, layout-dependent tasks beyond free-form QA. We ask: *do the phases of a staged adaptation pipeline contribute monotonically, or does staging induce capability trade-offs whose best checkpoint differs by task?*

Financial document understanding is a compelling testbed for staged adaptation because the domain simultaneously demands regulatory knowledge, numerical reasoning, and language-specific competence, capabilities that naturally map to distinct training phases. Each country maintains its own disclosure regime (e.g. SEC 10-K filings in the US, FSA filings in Japan), and domain-specialised LLMs for these regimes remain under-explored, especially for non-English domains where data is scarce and efficient training pipelines are valuable. Within this setting, Japanese financial disclosures are a particularly demanding case: their layout-rich pages of dense tables and multi-column statements must be read as images, and even state-of-the-art LLMs perform only marginally better than logistic regression on EDINET tasks (Sugiura et al., 2026). We study this through a three-phase *staged continual adaptation* pipeline—vision–language alignment, mathematical reasoning distillation, financial-domain adaptation—on an 8.4B-parameter model pairing SigLIP-v2 (Tschannen et al., 2025) with llm-jp-4-8B-Instruct (LLM-jp, 2025a;b) via a two-layer MLP projector (Zhang et al., 2023; Li et al., 2024). The un-fine-tuned instruction baseline serves as Phase 0; Phase 1 aligns visual and language representations using Japanese captioning/VQA corpora and OCR/visual QA synthesised from government PDFs (Cabinet Office, FSA, MOF); Phase 2 distils chain-of-thought mathematical reasoning from Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025b); Phase 3 injects financial-domain knowledge via text-only financial QA (Chen et al., 2021; 2022; Zhu et al., 2021) and visual QA from government PDFs. As testbeds we adopt EDINET disclosures (Sugiura et al., 2026) for real-world financial tasks, JP Fin Harness (Hirano, 2023) for Japanese finance knowledge, and GSM8K (Cobbe et al., 2021) for mathematical ability. All three are *text-input* evaluations: they probe how staging reshapes the backbone's financial and numerical capabilities, not the in-modality page-reading skill that the visual stages target, which

[1]Kyoto University, Kyoto, Japan. Correspondence to: Atsushi Yanagisawa <yanagisawa.atsushi.62t@st.kyoto-u.ac.jp>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

available benchmarks do not yet measure and we leave to future work. Within this scope, the phases interact non-additively, and a skip-phase ablation that omits Phase 2 isolates the resulting trade-off between numerical reasoning and domain priors.

Contributions: **(1)** evidence that the phases interact *non-additively* on EDINET and JP Fin Harness (Hirano, 2023): asymmetric forgetting (procedural skill survives, declarative knowledge is overwritten), repair rather than installation by the domain phase, and a suggestive cross-phase enabling effect, all anchored to the un-fine-tuned baseline; **(2)** a three-phase *staged continual adaptation* pipeline that trains end-to-end on page images, the visual stages serving as an adaptation vehicle whose in-modality skill our text-input benchmarks do not yet probe; **(3)** a reproducible synthetic-data pipeline generating OCR, visual QA, and domain-specific financial QA from government PDFs, with code and the resulting datasets released publicly;[1] **(4)** an XML-structured chain-of-thought distillation recipe from Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025b).

2. Model Design and Training Pipeline

2.1. Architecture

The model consists of three components: a vision encoder, an MLP projector, and a language model (overall architecture diagram in Figure 1, Appendix A). We adopt SigLIP-v2 SO400M-patch14-384 (Tschannen et al., 2025) as the vision encoder. Inspired by LLaVA-style projectors (Liu et al., 2023; Zhang et al., 2023; Li et al., 2024), the projector is a two-layer MLP (Linear → GELU → Linear) that maps visual features into the language model's embedding space and bridges the visual and linguistic representations. The language model is llm-jp-4-8B-Instruct (LLM-jp, 2025a) (8B parameters); together with the vision encoder (approximately 0.4B) the total parameter count is approximately 8.4B.

2.2. Data Generation from Government-Published PDFs

The synthetic supervision at the core of this study comes from two sources: publicly available PDFs issued by the Cabinet Office, FSA, and MOF, and public math corpora (see Table 3, Appendix A). We chose these agencies because they publish a large volume of finance-related materials.

From these government-published PDFs we construct three single-page datasets using *Qwen3-VL-32B-Instruct* (Qwen Team, 2025c) as the teacher: **(i) an OCR dataset** that targets text transcription of each page, following LLaVAR (Zhang et al., 2023); **(ii) an extractive/reasoning visual-QA dataset** in which the first turn extracts a specific element from the page and the second turn asks for the evidence grounding that answer, in the spirit of LLaVAR-2 (Zhou et al., 2025); and **(iii) a finance-specific QA dataset** that tests finance knowledge grounded in each page, with questions ranging from simple content extraction to multi-step numerical computations over the extracted content. For dataset (iii) only, we manually identify finance-relevant sub-sites within the Cabinet Office, FSA, and MOF web presences and restrict the source PDFs to those sub-sites. Examples and the full prompt templates are provided in Appendices B and C.

For the math reasoning dataset we draw problems from six publicly available math corpora—MGSM-ja (SB Intuitions, 2024), MathInstruct (TIGER-Lab, 2023), OpenMathReasoning (NVIDIA, 2025), OpenR1-Math-220k (Open-R1, 2025), orca-math-word-problems-200k (Microsoft, 2024), and NuminaMath-1.5 (AI-MO, 2025)—and randomly sample 150K problems in total across these corpora. Each sampled problem is re-annotated by prompting Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025b) to emit its response as a restated problem, reasoning process, and final answer inside `<Problem>`, `<Thinking>`, and `<Answer>` tags—a format chosen to localise intermediate-reasoning errors and to fit the multi-step numerical reasoning common in financial documents.

2.3. Three-Stage Training

As shown in Table 1, training proceeds in three phases (P1, P2, P3), each comprising one or more stages (S). **Phase 1 (P1, vision–language alignment)** is required because the pretrained LLM has no prior exposure to image tokens from SigLIP-v2's embedding space: Stage 1 (P1-S1) trains only the projector on STAIR Captions (Yoshikawa et al., 2017) and OCR data derived from government PDFs, with the vision encoder and LLM frozen; Stage 2 (P1-S2) then updates the projector and the vision encoder and adapts the LLM with LoRA (Hu et al., 2022), using extractive and reasoning QA we constructed from government PDFs together with the JA-VG-VQA conversation dataset (LLM-jp, 2024) and SakanaAI JA-VG-VQA-500 (SakanaAI, 2024). **Phase 2 (P2, mathematical reasoning distillation)** is motivated by the observation that financial documents require multi-step numerical deduction (e.g., consistency checks across tables): we construct approximately 150K math problems

[1] Code: `https://github.com/AtsushiYanaigsawa768/Compass`. The synthesised datasets are released on the Hugging Face Hub (Appendix B).

Table 1. Training phases and modules updated at each stage (P = Phase; S = Stage). "Vision" = SigLIP-v2 encoder; "Projector" = 2-layer MLP; "LLM" = llm-jp-4-8B-Instruct. "LoRA" (Hu et al., 2022) denotes LLM adaptation via low-rank adapters; "QLoRA" (Dettmers et al., 2023) denotes LoRA on a 4-bit quantized base model, used at P2-S1.

Phase	Updated	Primary Objective
P1-S1	Projector	Alignment of image and language embeddings (vision and LLM frozen)
P1-S2	Vision + Projector + LLM (LoRA)	Visual instruction following in Japanese
P2-S1	LLM (QLoRA)	Chain-of-thought distillation for math reasoning
P3-S1	LLM (LoRA)	Text-only financial-domain adaptation
P3-S2	Projector + LLM (LoRA)	Visual domain QA on government financial documents

with XML-structured chain-of-thought supervision (full dataset details in Appendix A) and train the LLM (via QLoRA-SFT) using Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025b) as the teacher. **Phase 3 (P3, financial-domain adaptation)** re-introduces domain-specific supervision, as general-purpose math distillation does not cover financial terminology or reporting conventions: P3-S1 applies LoRA to the LLM on text-only financial benchmarks (FinQA, ConvFinQA, TAT-QA) (Chen et al., 2021; 2022; Zhu et al., 2021), while P3-S2 trains the projector together with LoRA adapters on the LLM using visual domain QA from government PDFs; the vision encoder is frozen throughout P3. Hyperparameters, training-loss curves, and an example model response on a finance-related reasoning problem are provided in Appendix A; the full prompt templates used for synthetic data generation are listed in Appendix C.

3. Experimental Results

Per-phase evaluation. Table 2 reports scores on five benchmarks for the un-fine-tuned baseline (P0), after each training phase (P1–P3), and for the skip-P2 ablation (P1→P3). P0→P1 raises JP Fin Harness (45.0→53.5), GSM8K (58.5→64.3), EDINET fraud (0.516→0.554), and EDINET industry (4.7→8.7). From P1, GSM8K rises to 73.8 (P2) before slipping to 71.6 (P3), and EDINET fraud peaks at P3 (0.614), while JP Fin Harness (53.5→31.6→49.6) and EDINET industry (8.7→7.9→8.2) decline past P1. (EDINET earnings is near-chance throughout and is not interpreted.) These are single greedy-decoding runs; the

Table 2. Evaluation against the un-fine-tuned baseline (P0), after each training phase (P1–P3), and for the skip-P2 ablation (P1→P3). † ROC-AUC; other metrics are Accuracy. Bold marks the best score per row. EDINET earnings is near-chance for all our checkpoints (ROC-AUC ≈0.46–0.52), so the bold entry in that row is the numerical maximum, not a meaningful capability difference. Scores are single greedy runs; see Table 9 for multi-seed means.

Benchmark	P0	P1	P2	P3	P1→P3
GSM8K	58.5	64.3	**73.8**	71.6	59.6
EDINET earnings†	**0.524**	0.457	0.500	0.488	0.491
EDINET fraud†	0.516	0.554	0.534	**0.614**	0.499
EDINET industry	4.7	8.7	7.9	8.2	**10.4**
JP Fin Harness	45.0	53.5	31.6	49.6	**56.2**

3-seed means in Table 9 are lower and re-order some rows (e.g. GSM8K P2 69.4 ± 1.4, and the EDINET fraud peak moves from P3 to P1 at 0.562 ± 0.008), so individual peaks should be read cautiously.

Skip-P2 ablation. Skipping Phase 2 (P1→P3) reveals a clear trade-off: mathematical reasoning drops sharply (GSM8K 59.6 vs. 71.6) and EDINET fraud falls (0.499 vs. 0.614), yet domain knowledge is better preserved, with JP Fin Harness and EDINET industry reaching their overall peaks (56.2 and 10.4). Phase 2 thus installs numerical reasoning but erodes domain priors that Phase 3 alone cannot fully restore.

External-model context. For perspective on absolute quality, we compare our checkpoints against larger external models: Llama-3.3-70B, GPT-5, and DeepSeek-R1 on EDINET, and the official JP Fin Harness 0-shot leaderboard (Hirano, 2023). The full tables are deferred to Appendix A.6. In its best single run, our P3 reaches 0.614 ROC-AUC on EDINET fraud, exceeding the much larger Llama-3.3-70B (0.590) and DeepSeek-R1 (0.56) and approaching GPT-5 (0.62) under a matched text-input setting. This parity rests on one run: the 3-seed fraud mean is lower (0.547 ± 0.005; Table 9). A well-adapted small model can thus rival far larger ones on specific financial tasks, though ours trails every external baseline on the harder EDINET industry task.

Per-task JP Fin Harness breakdown. Subtask behaviour is heterogeneous (Table 5, Appendix A): the P0→P1 gain concentrates on `chabsa` (0.859→0.943), `cma_basics`, and `security_sales_1`; every subtask collapses at P2 and only partly recovers at P3; and skip-P2 (P1→P3) takes the best score on four of five subtasks, while `cpa_audit` alone peaks at P0 (0.276). The `chabsa` collapse and partial re-

covery (0.943→0.581→0.935) mirrors the aggregate declarative-knowledge trajectory discussed below.

4. Discussion

Beyond recency. A natural reading of Table 2 is *recency*—each benchmark peaks just after the phase that supervises it—yet this holds for only two of our signals (GSM8K at P2; single-run EDINET fraud at P3). JP Fin Harness and EDINET industry instead peak at the skip-P2 ablation, not at the financial phase P3, and both sit below their P1 values by P3; EDINET earnings stays near-chance throughout (ROC-AUC 0.46–0.52), so we do not read its P0 maximum as a capability. That the best checkpoint is usually *not* the most recently supervised one motivates examining how the phases interact.

Forgetting is asymmetric. The interference between phases is direction-dependent. Procedural math skill installed at P2 largely survives later financial tuning (GSM8K 73.8→71.6 at P2→P3, and stable across seeds), whereas the declarative financial knowledge acquired at P1 is overwritten by the math phase (JP Fin Harness 53.5→31.6 at P1→P2; evidence-grounded `chabsa` collapses likewise). These two effects come from different perturbations, and the math corpus is far larger than the financial ones, so the asymmetry is partly entangled with phase ordering and corpus size. We therefore read "procedural skills resist later tuning while declarative knowledge does not" as a hypothesis about *kinds* of capability within this pipeline rather than an established law; our evidence is limited to financial declarative knowledge (`chabsa` is financial sentiment), and a non-financial Japanese language task would be needed to test a broader linguistic claim.

The domain phase repairs more than it installs. Most financial ability is acquired already at the general alignment phase P1, destroyed at P2, and only partly restored at P3, which never regains the P1 level. The same trajectory appears on two independent benchmarks (JP Fin Harness and EDINET industry). The dedicated financial phase thus performs damage *repair* rather than acquisition, and skipping the reasoning phase (skip-P2) is best on both because nothing needs repairing. Why P3 falls short of P1 is, however, confounded: overwritten knowledge may be intrinsically hard to recover, or the P3 corpus is simply far smaller than P1's; a volume-matched or interleaved P3 would separate these.

Compositional enabling and design implications. One effect hints at positive composition rather than mere interference: removing the math phase drops EDINET fraud below the untrained baseline (skip-P2 0.499 vs. P0 0.516), as if reasoning distillation were a prerequisite for the later phase to express fraud detection. The effect is small, though: a 0.017 ROC-AUC gap astride 0.5 on a roughly 200-example test set, and under multiple seeds the fraud peak moves to P1 (Table 9), so the ordering is not robust. We therefore report it as a hypothesis and flag a multi-seed skip-P2 ablation as the priority follow-up. Taken together, a later phase can erase, preserve, or enable what an earlier one produced, so per-phase contributions are not decomposable and no single checkpoint is best for every task. The effective levers are then *scheduling and parameter isolation*: per-phase adapters composed rather than overwritten, replay or interleaving, and phase reordering, rather than more stages or data. This aligns with the workshop's sustainability aims.

Limitations and scope. (i) We provide multi-seed checks in Appendix A.8 (Table 9), which are broadly stable for GSM8K and earnings; the skip-P2 ablation could not be included for resource reasons, making it—together with the fraud single-run variance noted above—the clearest target for future work. (ii) The results characterise a single model (8.4B), language (Japanese), and domain (finance), so we read the phase interactions as properties of this specific pipeline. (iii) Training is restricted to single-page inputs; multi-page training is a natural follow-up. (iv) P2 uses QLoRA while other stages use full-precision LoRA, so a small contribution from quantization cannot be excluded.

5. Conclusion

We tracked an 8.4B-parameter multimodal model for Japanese financial documents from an un-fine-tuned baseline through three adaptation phases. The phases interact non-additively: math-only distillation overwrites declarative domain knowledge while procedural skill persists, and the domain phase repairs more than it installs. Per-phase contributions are thus not decomposable and the best checkpoint is task-dependent, so the key design levers are scheduling and parameter isolation rather than additional stages. Though scoped to Japanese finance, the methodology should transfer to other financial-disclosure regimes where specialised LLMs are needed.

Acknowledgements

This work was supported by the "Development Acceleration Use" program of ABCI 3.0, which is pro-

vided by AIST and AIST Solutions. This work was also supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

AI-MO. NuminaMath-1.5 (huggingface dataset). https://huggingface.co/datasets/AI-MO/NuminaMath-1.5, 2025.

Anthropic. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., and Wang, W. Y. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., and Wang, W. Y. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, 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., Canton Ferrer, 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., Arrieta Ibarra, I., 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., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hirano, M. Japanese language model financial evaluation harness. https://github.com/pfnet-research/japanese-lm-fin-harness, 2023. Accessed: 2026-04-12.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Kakimoto, G. and Yanagisawa, A. Compass: Domain-specific financial QA dataset (huggingface dataset). https://huggingface.co/datasets/Yana/ft-llm-2026-domain-specific-qa, 2026a.

Kakimoto, G. and Yanagisawa, A. Compass: OCR dataset from government-published Japanese PDFs (huggingface dataset). https://huggingface.co/datasets/Yana/ft-llm-2026-ocr-dataset, 2026b.

Kakimoto, G. and Yanagisawa, A. Compass: Extractive and reasoning visual-QA dataset (huggingface dataset). https://huggingface.co/datasets/Yana/ft-llm-2026-qa-dataset, 2026c.

Kakimoto, G. and Yanagisawa, A. Compass: XML-structured reasoning SFT dataset (huggingface dataset). https://huggingface.co/datasets/Yana/ft-llm-2026-reasoning-sft, 2026d.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. Oral; arXiv:2304.08485.

LLM-jp. ja-vg-vqa-conversation dataset. https://huggingface.co/datasets/llm-jp/ja-vg-vqa-conversation, 2024. Accessed: 2026-04-11.

LLM-jp. llm-jp-4-8b-instruct. https://huggingface.co/llm-jp/llm-jp-4-8b-instruct, 2025a. Hugging Face model card.

LLM-jp. LLM-jp organization page. https://huggingface.co/llm-jp, 2025b.

Meta. Llama-3.3-70B-Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, 2024. Hugging Face model card.

Microsoft. orca-math-word-problems-200k (huggingface dataset). https://huggingface.co/datasets/microsoft/orca-math-word-problems-200k, 2024. Companion dataset for Orca-Math (Mitra et al., arXiv:2402.14830).

NVIDIA. OpenMathReasoning (huggingface dataset). https://huggingface.co/datasets/nvidia/OpenMathReasoning, 2025.

Open-R1. OpenR1-Math-220k (huggingface dataset). https://huggingface.co/datasets/open-r1/OpenR1-Math-220k, 2025.

Qwen Team. Qwen-14B-Chat. https://huggingface.co/Qwen/Qwen-14B-Chat, 2023. Hugging Face model card.

Qwen Team. Qwen2-72B-Instruct. https://huggingface.co/Qwen/Qwen2-72B-Instruct, 2024. Hugging Face model card.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qwen Team. Qwen3-30B-A3B-Thinking-2507. https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507, 2025b. Hugging Face model card.

Qwen Team. Qwen3-VL-32B-Instruct. https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct, 2025c. Hugging Face model card.

SakanaAI. JA-VG-VQA-500 dataset. https://huggingface.co/datasets/SakanaAI/JA-VG-VQA-500, 2024. Accessed: 2026-04-12.

SB Intuitions. MGSM_ja (huggingface dataset). https://huggingface.co/datasets/sbintuitions/MGSM_ja, 2024. Japanese subset of the Multilingual Grade School Math benchmark.

Sugiura, I., Ishida, T., Makino, T., Tazuke, C., Nakagawa, T., Nakago, K., and Ha, D. EDINET-Bench: Evaluating LLMs on complex financial tasks using Japanese financial statements. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.08762.

TIGER-Lab. MathInstruct (huggingface dataset). https://huggingface.co/datasets/TIGER-Lab/MathInstruct, 2023. Instruction-tuning corpus accompanying MAmmoTH.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., and Zhai, X. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. STAIR Captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 417–421, Vancouver, Canada, 2017. Association for Computational Linguistics.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

Zhou, S., Zhang, R., Zhou, Y., and Chen, C. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025. Introduces the LLaVAR-2 dataset; arXiv:2412.16364.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., and Chua, T.-S. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3277–3287, 2021.

A. Supplementary Tables, Figures, and Training Behavior

A.1. Model Architecture Diagram

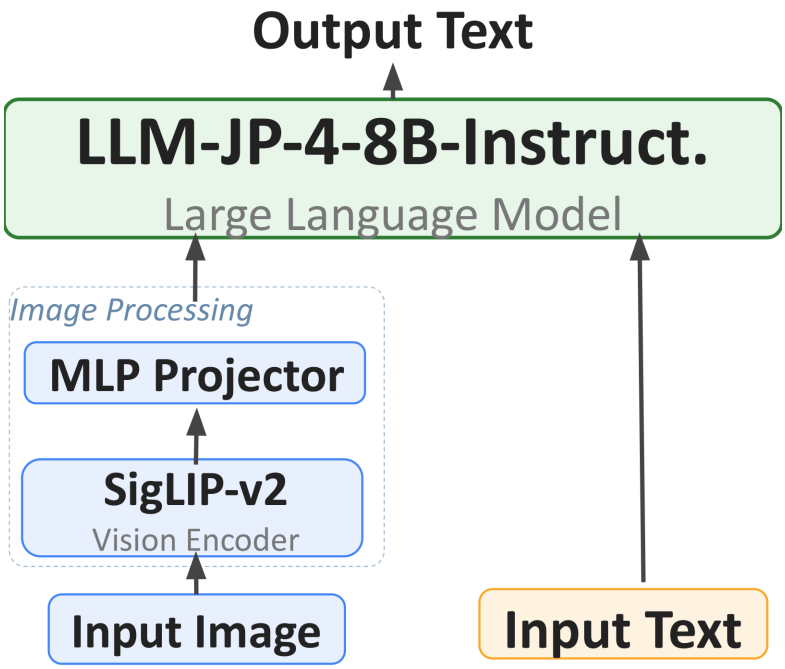


Figure 1. Architecture of the proposed model. The input image is connected to the language model through SigLIP-v2 and the MLP projector, and is processed jointly with the input text. By reading document pages directly as images, the model generates output text while preserving layout information such as tables, body text, and numerical arrangements.

A.2. Staged Learning Pipeline Diagram

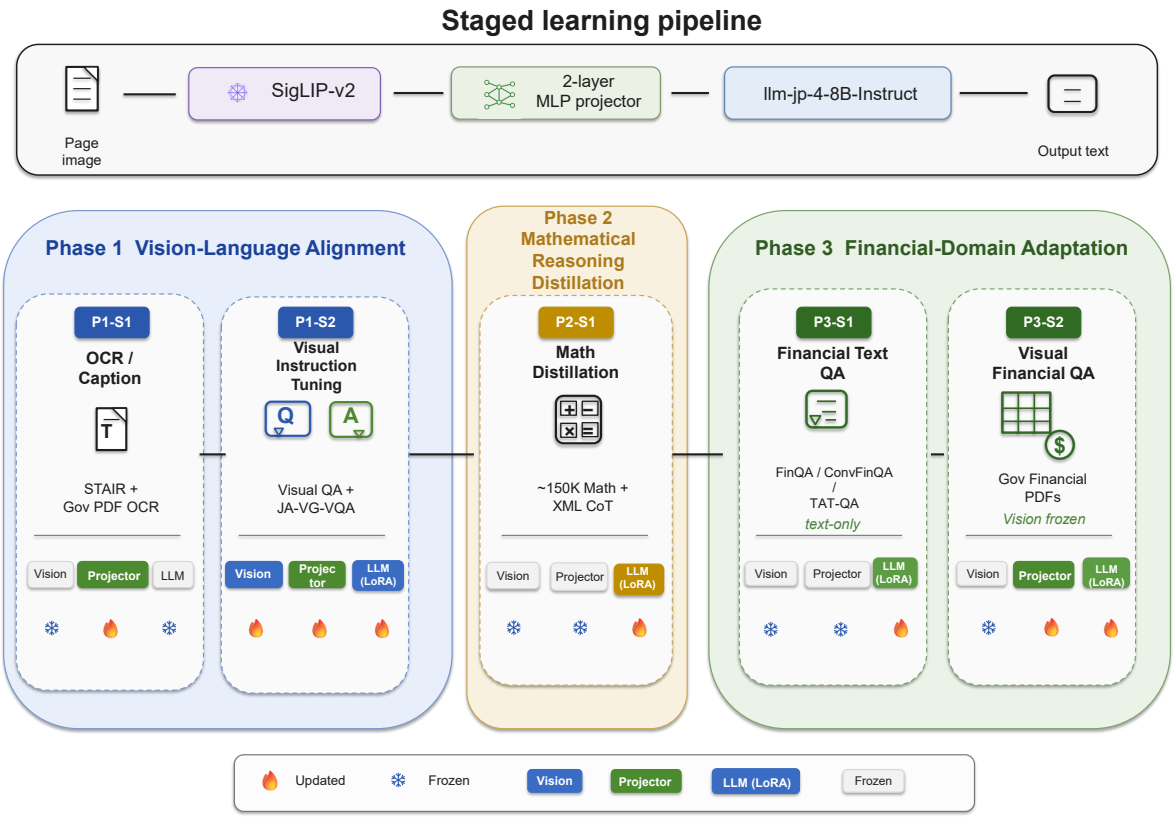


Figure 2. Overview of the three-phase staged continual adaptation pipeline. **Phase 1 (Vision-Language Alignment)**: P1-S1 trains only the projector on STAIR Captions and government-PDF OCR (vision encoder and LLM frozen); P1-S2 fine-tunes all three components (vision, projector, LLM via LoRA) on Visual QA and JA-VG-VQA. **Phase 2 (Mathematical Reasoning Distillation)**: P2-S1 updates the LLM via QLoRA on approximately 150K math problems with XML-structured chain-of-thought supervision from Qwen3-30B-A3B-Thinking-2507 (vision encoder and projector frozen). **Phase 3 (Financial-Domain Adaptation)**: P3-S1 adapts the LLM to text-only financial benchmarks (FinQA, ConvFinQA, TAT-QA); P3-S2 then trains the projector together with LoRA adapters on the LLM using visual financial QA from government PDFs (vision encoder frozen throughout P3). Flame icons indicate trainable modules; snowflake icons indicate frozen modules.

A.3. Dataset Overview

Table 3. Datasets used for training at each phase, with per-dataset sample counts and per-type subtotals. "Examples" counts captions for STAIR, image–QA pairs for VQA-style data, and problem–solution pairs for math/financial QA. Public-dataset counts are taken from the train splits of the original releases; the math corpus is a random sample of 150K problems drawn across the six listed sources (each entry shows the source's total available size). The datasets constructed in this work (synthesised from government PDFs) are publicly released on the Hugging Face Hub (Appendix B).

Phase	Type	Content	Examples	Type total
P1-S1	OCR / Caption	STAIR Captions (Yoshikawa et al., 2017)	820,310	1,084,952
		OCR data from government PDFs (synthesised)	264,642	
P1-S2	VQA	Extractive + reasoning QA from government PDFs (synthesised)	264,642	365,137
		JA-VG-VQA conversational (LLM-jp, 2024)	99,995	
		JA-VG-VQA-500 (SakanaAI, 2024)	500	
P2-S1	Reasoning SFT	sbintuitions/MGSM_ja (SB Intuitions, 2024)	250	∼150,000
		TIGER-Lab/MathInstruct (TIGER-Lab, 2023)	262,040	
		nvidia/OpenMathReasoning (NVIDIA, 2025)	306,000	
		open-r1/OpenR1-Math-220k (Open-R1, 2025)	220,000	
		microsoft/orca-math-word-problems-200k (Microsoft, 2024)	200,035	
		AI-MO/NuminaMath-1.5 (AI-MO, 2025)	896,215	
P3-S1	Financial QA	FinQA train (Chen et al., 2021)	6,251	22,503
		ConvFinQA train (Chen et al., 2022)	3,037	
		TAT-QA train (Zhu et al., 2021)	13,215	
P3-S2	Visual Financial QA	Domain-specific financial QA from government PDFs (Cabinet Office, FSA, MOF; synthesised)	32,484	32,484

A.4. Training Hyperparameters

Table 4. Main training hyperparameters by stage. "P1-S2" is the primary projector-and-vision-update sub-step; "P1-S2B (LLM/Vision)" denotes the subsequent step where the LLM is adapted with LoRA using per-module learning rates. "P2-S1 (SFT)" lists the hyperparameters of the chain-of-thought supervised distillation sub-phase; this stage uses QLoRA (4-bit NF4 quantized base with LoRA adapters), whereas all other stages use full-precision LoRA. P3-S2 shares the P3-S1 configuration unless otherwise noted.

Phase	LR	Batch	LoRA (r, α)	Epochs
P1-S1	1e-3	2	None	2
P1-S2	2e-5	2	64, 128	1
P1-S2B (LLM)	1e-5	2	64, 128	1
P1-S2B (Vision)	2e-6	2	64, 128	1
P2-S1 (SFT, QLoRA)	2e-4	2	32, 64	1
P3-S1 / P3-S2	5e-6	2	16, 32	3

A.5. Per-Task Results on JP Fin Harness

Table 5. Breakdown on JP Fin Harness tasks (Accuracy). P0 is the un-fine-tuned baseline; P3-1 / P3-2 are the two sub-stages of Phase 3; P1→P3 is the skip-P2 ablation. Bold indicates the best-performing model for each task.

JP Fin task	P0	P1	P2	P3-1	P3-2	P1→P3
chabsa	0.8591	0.9428	0.5814	0.9357	0.9353	**0.9479**
cma_basics	0.3421	0.5526	0.2105	0.3947	0.3947	**0.5789**
cpa_audit	**0.2764**	0.2236	0.1683	0.2487	0.2513	0.2312
fp2	0.2989	0.3432	0.1789	0.2863	0.3011	**0.3832**
security_sales_1	0.4737	0.6140	0.4386	0.5614	0.5965	**0.6667**
Average	0.4500	0.5352	0.3155	0.4854	0.4958	**0.5616**

A.6. External-Model Comparison on EDINET and JP Fin Harness

These tables place our checkpoints next to substantially larger models, as indicative context rather than head-to-head results: the Llama-3.3-70B row is evaluated under the same setup as ours, whereas the GPT-5 and DeepSeek-R1 numbers are taken from EDINET-Bench (Sugiura et al., 2026) and the JP Fin Harness rows from the official 0-shot leaderboard (Hirano, 2023), so they partly draw on different evaluation sources. We emphasise that the EDINET evaluation here is *text-only*: the financial statements (Summary, balance sheet, profit-and-loss, and cash-flow tables) are tokenised and fed to the LLM backbone as text, so the page-image modality that is the focus of the main pipeline is not exercised by this benchmark. To keep the comparison consistent, the GPT-5 and DeepSeek-R1 entries are read from the same input column of EDINET-Bench as our own configuration; we further restrict the comparison to LLM-based systems.

Table 6 places our P3 checkpoint alongside Llama-3.3-70B, GPT-5, and DeepSeek-R1 on EDINET, and Table 7 compares our JP Fin Harness results with the official 0-shot leaderboard (Hirano, 2023). On EDINET fraud, our best single run (P3, 0.614 ROC-AUC) exceeds Llama-3.3-70B (0.590, same setup as ours) and the DeepSeek-R1 number reported by EDINET-Bench (Sugiura et al., 2026) (0.56) and approaches GPT-5 (0.62); under three seeds, however, the fraud mean is lower (0.547 ± 0.005) and peaks at P1, so we read the parity with GPT-5 as a best-run result rather than a stable one. On EDINET earnings every system except GPT-5 is at or below chance (ROC-AUC $\leq$0.5: ours 0.488, Llama-3.3-70B 0.410, DeepSeek-R1 0.46; even GPT-5 reaches only 0.62), so this row is not informative and we draw no conclusions from it. EDINET industry remains a weakness: our P3 result (8.2) and even our best checkpoint (10.4, P1→P3; Table 2) trail every external baseline (14.0–31.0). On JP Fin Harness, evaluated 0-shot with the harness's standard prompting to match the leaderboard, our P1→P3 ablation (56.2) exceeds Qwen-14B-Chat (49.1) and Meta-Llama-3-8B-Instruct (44.7) but trails Qwen2-72B-Instruct (67.7) and Claude 3.5 Sonnet (77.0). On GSM8K (Table 8), our P2 (73.8) exceeds Qwen-14B Chat (59.3) but remains 11–16 points below the base-model Llama 3 8B (84.5) and Qwen3 8B-base (89.84).

Table 6. EDINET comparison. All scores use the EDINET-Bench text-input setting in which the financial statements (Summary, balance sheet, profit-and-loss, and cash-flow tables) are supplied to the LLM as text rather than page images; the vision modality is not exercised by this benchmark. Llama-3.3-70B (Meta, 2024) is evaluated under the same setup as ours; the GPT-5 and DeepSeek-R1 numbers are the corresponding column from EDINET-Bench (Sugiura et al., 2026) and are listed for context only, since cross-source comparisons are indicative rather than head-to-head. † ROC-AUC; *industry* is Accuracy. Bold marks the best score per row.

Benchmark	Llama-3.3-70B	GPT-5	DeepSeek-R1	Ours (P3)
EDINET earnings†	0.410	**0.62**	0.46	0.488
EDINET fraud†	0.590	**0.62**	0.56	0.614
EDINET industry	14.0	**31.0**	16.0	8.2

Table 7. JP Fin Harness comparison. External scores are from the official 0-shot leaderboard (Hirano, 2023). Ours (P0) is the un-fine-tuned llm-jp-4-8B-Instruct baseline; we additionally show Ours (P1), Ours (P1→P3), and Ours (P3) because JP Fin Harness peaks at P1→P3. Bold: best score.

Model	Size	Avg.
Claude 3.5 Sonnet (Anthropic, 2024)	proprietary	**77.0**
Qwen2-72B-Instruct (Qwen Team, 2024)	72B	67.7
Qwen-14B-Chat (Qwen Team, 2023)	14B	49.1
Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024)	8B	44.7
Ours (P0)	8B	45.0
Ours (P1)	8B	53.5
Ours (P1→P3)	8B	56.2
Ours (P3)	8B	49.6

A.7. GSM8K Comparison with External Models

Table 8. GSM8K accuracy comparison with external models and our four checkpoints (P0–P3). Qwen-14B Chat (Qwen Team, 2023) was evaluated in our setting; Qwen3 8B-base (Qwen Team, 2025a) and Llama 3 8B (Grattafiori et al., 2024) results are taken from their respective technical reports (both base-model values). Ours (P0) is the un-fine-tuned llm-jp-4-8B-Instruct baseline. Our Phase 2/3 checkpoints are evaluated on the GSM8K test split after distillation on a 150K-problem mix randomly sampled from MathInstruct, OpenMathReasoning, OpenR1-Math-220k, orca-math-word-problems-200k, NuminaMath-1.5, and MGSM-ja; some of these aggregate corpora include GSM8K-style problems, so the comparison should be read with this in-distribution fine-tuning in mind. Where GSM8K-derived problems appear in the training mix, they originate solely from the GSM8K *train* split; the GSM8K *test* split used here for evaluation is held out and never enters P2 distillation, so the reported gains are not attributable to test-set leakage. Bold indicates the highest score.

Benchmark	Qwen-14B Chat	Qwen3 8B-base	Llama 3 8B	Ours (P0)	Ours (P1)	Ours (P2)	Ours (P3)
GSM8K	59.3	**89.84**	84.5	58.5	64.3	73.8	71.6

A.8. Multi-Seed Robustness Evaluation

Table 9. Multi-seed robustness evaluation (temperature 0.7, top_p 0.95, 3 seeds). Values are mean ± standard deviation. GSM8K and JP Fin Harness subtasks are Accuracy; EDINET metrics are ROC-AUC (†) or Accuracy.

Benchmark	P1	P2	P3
GSM8K	62.6 ± 0.5	69.4 ± 1.4	68.8 ± 1.3
EDINET earnings†	0.475 ± 0.014	0.484 ± 0.025	0.473 ± 0.025
EDINET fraud†	0.562 ± 0.008	0.523 ± 0.008	0.547 ± 0.005
EDINET industry	6.7 ± 0.4	6.8 ± 0.3	6.0 ± 0.4
JP Fin chabsa	0.909 ± 0.001	0.587 ± 0.001	0.924 ± 0.001
JP Fin cma_basics	0.500 ± 0.021	0.219 ± 0.012	0.360 ± 0.106
JP Fin cpa_audit	0.188 ± 0.013	0.168 ± 0.000	0.187 ± 0.020
JP Fin fp2	0.313 ± 0.006	0.182 ± 0.003	0.310 ± 0.019
JP Fin security_sales_1	0.602 ± 0.036	0.480 ± 0.030	0.561 ± 0.035

A.9. Example Model Response to a Finance-Related Problem

Input

```
Given that revenue was 100 in the previous period and 120 in the current period,
compute the year-over-year growth rate.
```

Output

```
<Problem>Given that revenue was 100 in the previous period and 120 in the current
period, compute the year-over-year growth rate.</Problem>

<Thinking>
To compute the year-over-year growth rate, I need to use the formula:

Growth Rate = (Current Value - Previous Value) / Previous Value * 100

In this case:
- Previous Value = 100

So, the calculation would be:
(120 - 100) / 100 * 100

Let me simplify this:
(20 / 100) * 100 = 20%

Therefore, the year-over-year growth rate is 20%.
</Thinking>

<Answer>
20
</Answer>
```

Figure 3. Example model response when the input is a finance-related reasoning problem. The model emits the input restated as `<Problem>`, intermediate reasoning as `<Thinking>`, and the final boxed answer as `<Answer>`—the structure that Phase 2 distillation installs.

A.10. Training Loss Curves

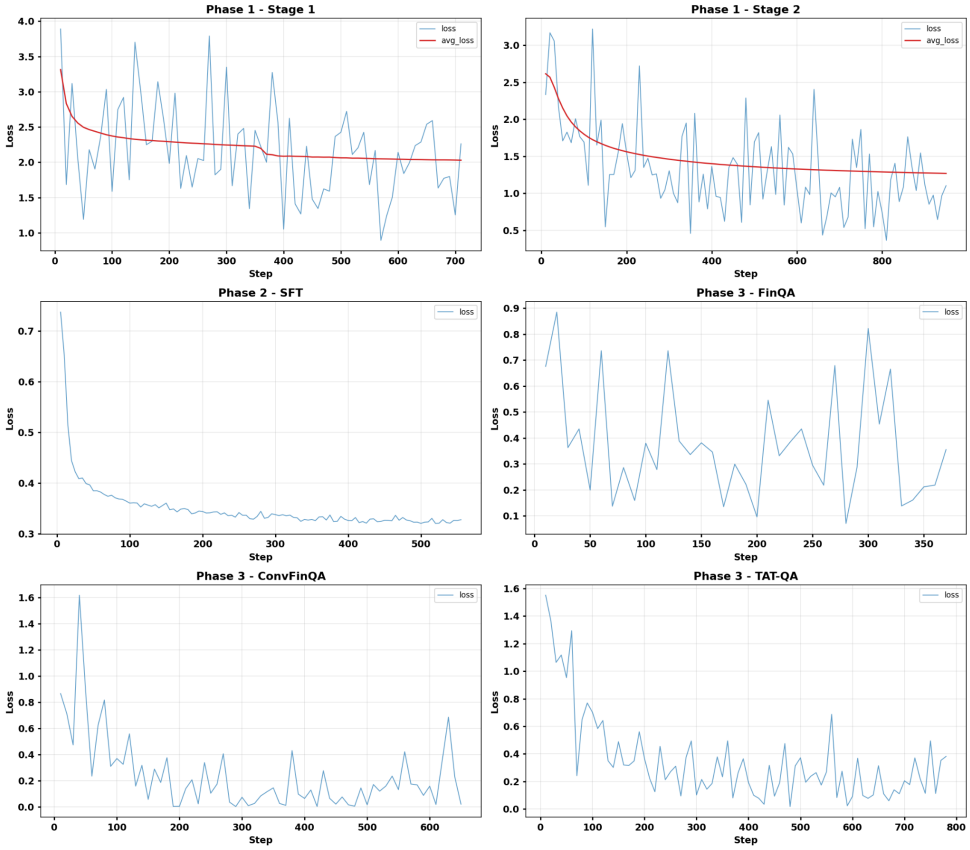


Figure 4. Loss curves during three-stage training. Phase 1 achieves vision–language alignment, Phase 2 develops reasoning, and Phase 3 performs financial adaptation in sequence.

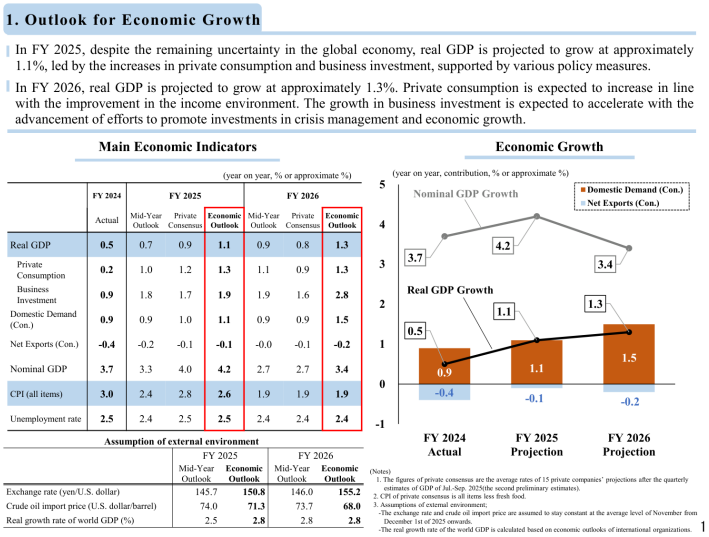
The loss curves in Figure 4 show stable decreases across Phase 1 (Stages 1 and 2), Phase 2 SFT, and Phase 3 (S1; the text-only financial stages shown here).

B. Example Training Data from Government PDFs

The datasets synthesised from government-published PDFs are publicly available on the Hugging Face Hub: the OCR corpus (Kakimoto & Yanagisawa, 2026b), the extractive and reasoning visual-QA corpus (Kakimoto & Yanagisawa, 2026c), and the domain-specific financial-QA corpus (Kakimoto & Yanagisawa, 2026a); the XML-structured reasoning-SFT corpus distilled at Phase 2 (Kakimoto & Yanagisawa, 2026d) is released as well. Representative examples follow.

B.1. OCR Example

Retrieved from `https://www5.cao.go.jp/keizai1/summary_fy2026.pdf` (page 2).



1. Outlook for Economic Growth

In FY 2025, despite the remaining uncertainty in the global economy, real GDP is projected to grow at approximately 1.1%, led by the increases in private consumption and business investment, supported by various policy measures.

In FY 2026, real GDP is projected to grow at approximately 1.3%. Private consumption is expected to increase in line with the improvement in the income environment. The growth in business investment is expected to accelerate with the advancement of efforts to promote investments in crisis management and economic growth.

Main Economic Indicators

(year on year, % or approximate %)

	FY 2024	FY 2025			FY 2026		
	Actual	Mid-Year Outlook	Private Consensus	Economic Outlook	Mid-Year Outlook	Private Consensus	Economic Outlook
Real GDP	**0.5**	0.7	0.9	**1.1**	0.9	0.8	**1.3**
Private Consumption	**0.2**	1.0	1.2	**1.3**	1.1	0.9	**1.3**
Business Investment	**0.9**	1.8	1.7	**1.9**	1.9	1.6	**2.8**
Domestic Demand (Con.)	**0.9**	0.9	1.0	**1.1**	0.9	0.9	**1.5**
Net Exports (Con.)	**-0.4**	-0.2	-0.1	**-0.1**	-0.0	-0.1	**-0.2**
Nominal GDP	**3.7**	3.3	4.0	**4.2**	2.7	2.7	**3.4**
CPI (all items)	**3.0**	2.4	2.8	**2.6**	1.9	1.9	**1.9**
Unemployment rate	**2.5**	2.4	2.5	**2.5**	2.4	2.4	**2.4**

Assumption of external environment

	FY 2025		FY 2026	
	Mid-Year Outlook	Economic Outlook	Mid-Year Outlook	Economic Outlook
Exchange rate (yen/U.S. dollar)	145.7	**150.8**	146.0	**155.2**
Crude oil import price (U.S. dollar/barrel)	74.0	**71.3**	73.7	**68.0**
Real growth rate of world GDP (%)	2.5	**2.8**	2.8	**2.8**

Economic Growth

(Notes)
1. The figures of private consensus are the average rates of 15 private companies' projections after the quarterly estimates of GDP of Jul.-Sep. 2025(the second preliminary estimates).
2. CPI of private consensus is all items less fresh food.
3. Assumptions of external environment;
-The exchange rate and crude oil import price are assumed to stay constant at the average level of November from December 1st of 2025 onwards.
-The real growth rate of the world GDP is calculated based on economic outlooks of international organizations.

1

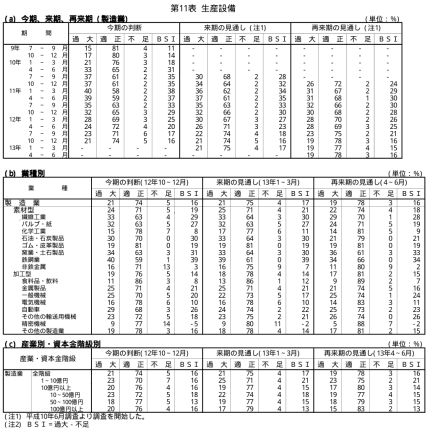
Figure 5. Source page.

```
1. Outlook for Economic Growth
In FY 2025, despite the remaining uncertainty in the global economy, real GDP is projected to
grow at approximately 1.1%, led by the increases in private consumption and business investment,
supported by various policy measures.
In FY 2026, real GDP is projected to grow at approximately 1.3%. Private consumption is expected
to accelerate with the improvement in the income environment. The growth in business investment
is expected to continue with the advancement of efforts to promote investments in crisis
management and economic growth.
Main Economic Indicators (year on year, %): rows include Real GDP, Private Consumption, Private
Residential Investment, Business Investment, Government Consumption, Public Investment, Nominal
GDP, CPI (all items); columns cover FY 2024 Actual and FY 2025 / FY 2026 under Mid-Year Outlook,
Private Consensus, and Economic Outlook scenarios.
Assumption of external environment: Exchange rate (yen/U.S. dollar), crude oil price (Dubai, U.S.
dollar/barrel), and real growth rate of world GDP (%) for FY 2025 and FY 2026.
[… transcript continues below this excerpt]
```

Figure 6. OCR transcript.

B.2. Visual QA Example

Retrieved from https://www.esri.cao.go.jp/jp/stat/hojin/doko/h12doko/00122hojin.pdf (page 12).



第11表 生産設備

(a) 今期、来期、再来期（製造業）　（単位：%）

期	今期の判断					来期の見通し（注1）					再来期の見通し（注1）				
	過大	適正	不足	BSI		過大	適正	不足	BSI		過大	適正	不足	BSI	
9年 7 － 9 月	15	81	4	11		-	-	-	-		-	-	-	-	
10 － 12 月	17	80	3	14		-	-	-	-		-	-	-	-	
10年 1 － 3 月	21	76	3	18		-	-	-	-		-	-	-	-	
4 － 6 月	33	65	2	31		-	-	-	-		-	-	-	-	
7 － 9 月	37	61	2	35		30	68	2	28		-	-	-	-	
10 － 12 月	37	61	2	35		34	64	2	32		26	72	2	24	
11年 1 － 3 月	40	58	2	38		36	62	2	34		31	67	2	29	
4 － 6 月	39	59	2	37		37	61	2	35		31	68	1	30	
7 － 9 月	35	63	2	33		35	63	2	33		32	66	2	30	
10 － 12 月	32	65	3	29		32	66	2	30		30	68	2	28	
12年 1 － 3 月	28	69	3	25		30	67	3	27		28	70	2	26	
4 － 6 月	24	72	4	20		26	71	3	23		28	69	3	25	
7 － 9 月	23	71	6	17		22	74	4	18		23	75	2	21	
10 － 12 月	21	74	5	16		21	74	5	16		19	78	3	16	
13年 1 － 3 月	-	-	-	-		21	75	4	17		19	77	4	15	
4 － 6 月	-	-	-	-		-	-	-	-		19	78	3	16	

(b) 業種別　（単位：%）

業種	今期の判断（12年10－12月）				来期の見通し（13年1－3月）				再来期の見通し（13年4－6月）			
	過大	適正	不足	BSI	過大	適正	不足	BSI	過大	適正	不足	BSI
製造業	21	74	5	16	21	75	4	17	19	78	3	16
素材型	24	71	5	19	25	71	4	21	22	74	4	18
繊維工業	33	63	4	29	33	64	3	30	29	70	1	28
パルプ・紙	32	63	5	27	32	63	5	27	24	71	5	19
化学工業	15	78	7	8	17	77	6	11	14	81	5	9
石油・石炭製品	30	70	0	30	33	64	3	30	21	79	0	21
ゴム・皮革製品	19	81	0	19	19	81	0	19	19	81	0	19
窯業・土石製品	34	63	3	31	33	64	3	30	36	61	3	33
鉄鋼業	40	59	1	39	39	61	0	39	34	66	0	34
非鉄金属	16	71	13	3	16	75	9	7	11	80	9	2
加工型	19	76	5	14	18	78	4	14	17	81	2	15
食料品・飲料	11	86	3	8	13	86	1	12	9	89	2	7
金属製品	25	71	4	21	25	71	4	21	21	74	5	16
一般機械	25	70	5	20	22	73	5	17	25	74	1	24
電気機械	16	78	6	10	16	78	6	10	14	83	3	11
自動車	29	68	3	26	24	74	2	22	25	73	2	23
その他の輸送用機械	23	72	5	18	23	75	2	21	26	74	0	26
精密機械	9	77	14	-5	9	80	11	-2	5	88	7	-2
その他の製造業	19	78	3	16	18	78	4	14	17	81	2	15

(c) 産業別・資本金階級別　（単位：%）

産業・資本金階級		今期の判断（12年10－12月）				来期の見通し（13年1－3月）				再来期の見通し（13年4－6月）			
		過大	適正	不足	BSI	過大	適正	不足	BSI	過大	適正	不足	BSI
製造業	全階級	21	74	5	16	21	75	4	17	19	78	3	16
	1－10億円	23	70	7	16	25	71	4	21	23	75	2	21
	10億円以上	20	76	4	16	19	77	4	15	17	80	3	14
	10－50億円	23	72	5	18	22	74	4	18	19	77	4	15
	50－100億円	18	77	5	13	19	77	4	15	18	79	3	15
	100億円以上	20	76	4	16	17	79	4	13	15	83	2	13

（注1）平成10年6月調査より調査を開始した。
（注2）BSI－過大－不足

Figure 7. Source page.

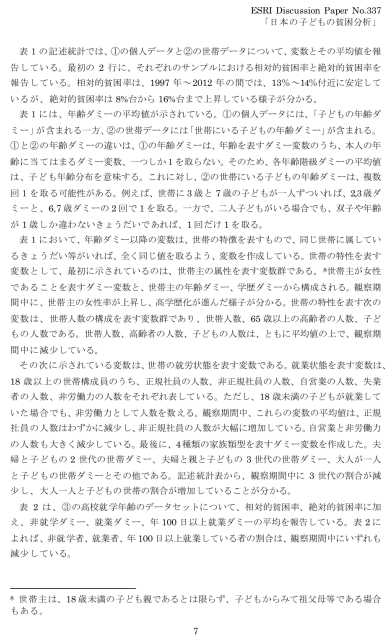
[Japanese (original)]
Extractive QA
Q: 表 (b) の「製造業」の「今期の判断 (12 年 10〜12 月)」の「過大」の値は何ですか？
A: 21
Reasoning QA
Q: 表 (b) の「製造業」の「今期の判断 (12 年 10〜12 月)」の「過大」の値は，どのセルから読み取られたか？
A: 表 (b) の「製造業」の行と「今期の判断 (12 年 10〜12 月)」の列，そして「過大」の交差するセルに「21」と記載されているため，そこから読み取られた。

[English translation]
Extractive QA
Q: In Table (b), what is the "excessive" value for "Manufacturing" under "Current assessment (Oct–Dec 2012)"?
A: 21
Reasoning QA
Q: From which cell was the "excessive" value for "Manufacturing" under "Current assessment (Oct–Dec 2012)" in Table (b) read?
A: It was read from the cell at the intersection of the "Manufacturing" row, the "Current assessment (Oct–Dec 2012)" column group, and the "excessive" sub-column in Table (b), where the value "21" is recorded.

Figure 8. Visual QA pair.

B.3. Domain-Specific Financial QA Example

Retrieved from https://www.esri.cao.go.jp/jp/esri/archive/e_dis/e_dis337/e_dis337.pdf (page 9).



ESRI Discussion Paper No.337
「日本の子どもの貧困分析」

表 1 の記述統計では、①の個人データと②の世帯データについて、変数とその平均値を報告している。最初の 2 行に、それぞれのサンプルにおける相対的貧困率と絶対的貧困率を報告している。相対的貧困率は、1997 年〜2012 年の間では、13%〜14%付近に安定しているが、絶対的貧困率は 8%台から 16%台まで上昇している様子が分かる。

表 1 には、年齢ダミーの平均値が示されている。①の個人データには、「子どもの年齢ダミー」が含まれる一方、②の世帯データには「世帯にいる子どもの年齢ダミー」が含まれる。①と②の年齢ダミーの違いは、①の年齢ダミーは、年齢を表すダミー変数のうち、本人の年齢に当てはまるダミー変数、一つしか 1 を取らない。そのため、各年齢階級ダミーの平均値は、子ども年齢分布を意味する。これに対し、②の世帯にいる子どもの年齢ダミーは、複数回 1 を取る可能性がある。例えば、世帯に 3 歳と 7 歳の子どもが一人ずついれば、2,3 歳ダミーと、6,7 歳ダミーの 2 回で 1 を取る。一方で、二人子どもがいる場合でも、双子や年齢が 1 歳しか違わないきょうだいであれば、1 回だけ 1 を取る。

表 1 において、年齢ダミー以降の変数は、世帯の特徴を表すもので、同じ世帯に属しているきょうだい等がいれば、全く同じ値を取るよう、変数を作成している。世帯の特性を表す変数として、最初に示されているのは、世帯主の属性を表す変数群である。*世帯主が女性であることを表すダミー変数と、世帯主の年齢ダミー、学歴ダミーから構成される。観察期間中に、世帯主の女性率が上昇し、高学歴化が進んだ様子が分かる。世帯の特性を表す次の変数は、世帯人数の構成を表す変数群であり、世帯人数、65 歳以上の高齢者の人数、子どもの人数である。世帯人数、高齢者の人数、子どもの人数は、ともに平均値の上で、観察期間中に減少している。

その次に示されている変数は、世帯の就労状態を表す変数である。就業状態を表す変数は、18 歳以上の世帯構成員のうち、正規社員の人数、非正規社員の人数、自営業の人数、失業者の人数、非労働力の人数をそれぞれ表している。ただし、18 歳未満の子どもが就業していた場合でも、非労働力として人数を数える。観察期間中、これらの変数の平均値は、正規社員の人数はわずかに減少し、非正規社員の人数が大幅に増加している。自営業と非労働力の人数も大きく減少している。最後に、4 種類の家族類型を表すダミー変数を作成した。夫婦と子どもの 2 世代の世帯ダミー、夫婦と親と子どもの 3 世代の世帯ダミー、大人が一人と子どもの世帯ダミーとその他である。記述統計表から、観察期間中に 3 世代の割合が減少し、大人一人と子どもの世帯の割合が増加していることが分かる。

表 2 は、③の高校就学年齢のデータセットについて、相対的貧困率、絶対的貧困率に加え、非就学ダミー、就業ダミー、年 100 日以上就業ダミーの平均を報告している。表 2 によれば、非就学者、就業者、年 100 日以上就業している者の割合は、観察期間中にいずれも減少している。

* 世帯主は、18 歳未満の子ども親であるとは限らず、子どもからみて祖父母等である場合もある。

7

Figure 9. Source page.

[Japanese (original)]
Q: 1997 年から 2012 年までの相対的貧困率の下限値は？（単位：%，小数点以下 0 桁）
A: 13

[English translation]
Q: What is the lower bound of the relative poverty rate from 1997 to 2012? (Unit: %, integer, no decimal places.)
A: 13

Figure 10. Domain-specific financial QA pair.

C. Prompt Templates for Synthetic Data Generation

This appendix lists the prompt templates used to generate OCR, visual QA, and financial-domain QA supervision from government-published PDFs. All prompts are reproduced below; minor redactions (noted in captions) remove auxiliary fields that were not reported as part of this paper's claims.

C.1. OCR Prompt

```
[SYSTEM]
You are a high-precision OCR assistant.
You will receive a single image as input.

Task:
- Extract only the text information written in the image as accurately
  and completely as possible.
- Transcribe the text exactly as written in the original script; do NOT
  translate or paraphrase.
- Consider the reading order of the text and arrange it in a natural
  reading sequence.
    * Imagine reading from top-left to bottom-right across the entire page.
    * Follow the order a human would read: headings -> body text ->
      figure/table captions -> footnotes/annotations.

Output Format:
- Output plain text only. Do not include explanations or summaries.
- Insert line breaks where lines or paragraphs change, preserving the
  original layout to some extent.
- Do not guess or infer unreadable or missing characters.

Prohibited Actions:
- Do not summarize or add explanatory text about the image content.
- Do not create new text that does not exist in the image.
- Do not translate or paraphrase the text.

[USER]
List all text that can be confirmed in the given image.
```

Listing 1. OCR generation prompt. Used to transcribe a single page image into plain text while preserving reading order.

C.2. Visual QA Prompts (Extractive and Reasoning)

```
[SYSTEM]
You are an AI visual assistant. You are viewing a single image.

Task:
Generate multiple single-turn QA pairs from the given document image.
Each QA should focus on different components within the image (text,
figures, tables, numbers, etc.) and should not overlap with each other.

Requirements for generated QA:
1. Only ask questions about content visible in the image that you can
   answer with confidence.
2. You may also answer when you can clearly determine that something
   does NOT exist in the image.
3. Do not include questions you cannot answer with confidence.
4. Ask questions about visual content (text, figures, tables, graphs,
   layout, etc.).

Question Diversity:
- Reading text / titles / headings
- Extracting numbers / dates / times
- Reading values from figures / graphs
- Understanding layout / structure
- Extracting data from tables

Output Format (JSON only, no additional text):
{
  "de_pairs": [
    {"question": "...", "answer": "..."},
    {"question": "...", "answer": "..."}
  ]
}

Notes:
- Answers must be based on information visible in the image.
- Do not ask about uncertain details.
```

```
- Each QA should focus on different parts of the image.
- Generate approximately 5-10 QA pairs.

[USER]
Generate multiple extractive QA pairs in JSON format, focusing on
different components within this document image.
```

Listing 2. Extractive QA-generation prompt. Produces multiple single-turn QA pairs grounded in visible content.

```
[SYSTEM]
You are an AI visual assistant. You are viewing a single image.

Task:
For a given reference QA (extractive QA), generate a corresponding
reasoning QA (why/where/how QA). The reasoning QA asks "why/where/how
was that extractive answer obtained?"

Rules for generating reasoning QA:

1. Generate a question that asks for the reason behind the reference
   QA's answer, or asks how the answer was derived.

2. When the reasoning process of the reference QA is very explicit
   (e.g., simple extraction):
   - The reasoning QA should focus on the source of the answer.
   - Example: "From which part of the image was this information
     extracted?"
   - "Where in the image is that value stated?"

3. When the reasoning process of the reference QA is complex:
   - The reasoning QA should infer relationships between related
     objects in the image.
   - Example: "What visual elements did you connect to reach this
     conclusion?"

4. Diversify the style of generated questions.
5. Answer the reasoning question using evidence from the image, but
   answer from the perspective of only viewing the image.
6. The reasoning QA must be closely related to the reference QA and
   the image. Avoid extending to new concepts.

Output Format (JSON only, no additional text):
{
  "dr": {"question": "...", "answer": "..."}
}
```

Listing 3. Reasoning QA-generation prompt. Given a reference extractive QA, produces a why/where/how follow-up.

C.3. Financial Domain QA Prompt

```
[SYSTEM]
You are an expert in financial and economic education.
Based on the attached document, create 3-5 question-answer pairs that
meet the following requirements.

Document Types:
Financial statements, earnings reports, economic reports, news
articles, etc. (one page).

Question Design Principles:

Purpose:
1. Document Comprehension: accurately extract necessary information
   from financial documents.
2. Financial Knowledge Acquisition: learn financial / economic
   concepts through problem solving.
3. Reasoning & Calculation: derive answers by combining multiple
   pieces of information.

Output Format (JSON only, no additional text):
{
  "qa_pairs": [
    {
      "question": "Clear, unambiguous question text.",
      "answer": "Concise answer with correct format / unit.",
      "reasoning": "Step-by-step reasoning process.",
      "explanation": "Brief explanation of the answer."
    }
```

```
  ]
}

Critical Rules:
- Generate BOTH question AND answer for each item.
- The answer must be accurate and derived from the document.
- Hallucination is strictly prohibited.
- Do not generate questions that cannot be answered from the document.
- Return an empty list if the image contains no financial or economic
  content.

[USER]
Analyze the attached financial document and create 3-5 question-answer
pairs following the specified format. For each question, provide the
correct answer with reasoning. Generate now.
```

Listing 4. Financial domain QA-generation prompt. Simplified presentation: auxiliary taxonomy fields used internally but not reported in this paper are omitted.