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 pairs vision–language alignment, reasoning distillation, and domain tuning. How each phase trades capabilities against the others, however, remains largely under-characterised. Tracking an 8.4B-parameter model from an un-fine-tuned baseline through three training phases on Japanese financial disclosures, we find that each benchmark peaks at a *different* phase. The best checkpoint is therefore task-dependent rather than simply the final one, with direct implications for checkpoint retention and compute budgeting under resource-constrained continual adaptation.

1. Introduction

Adapting a multimodal foundation model to a specialised domain rarely proceeds in a single step: contemporary recipes chain together vision–language alignment, reasoning distillation, and domain tuning, and 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 prediction 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?*

Japanese financial disclosures are a demanding case for such adaptation: 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 such as accounting-fraud detection and earnings forecasting (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 so each phase's contribution is measured against a common reference. Phase 1 aligns visual and language representations using existing Japanese captioning/VQA corpora together with OCR and visual QA we synthesise from PDFs published by Japanese government agencies—the Cabinet Office, the Financial Services Agency (FSA), and the Ministry of Finance (MOF); each visual-QA item pairs an extractive question that retrieves a specific element from the document image with a reasoning question that asks for the evidence grounding that answer. Phase 2 distils chain-of-thought from Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025); we include this phase because solving real-world financial tasks requires a non-trivial level of numerical reasoning. Phase 3 consumes text-only financial benchmarks (Chen et al., 2021; Zhu et al., 2021; Chen et al., 2022) together with visual QA we construct from financial documents to instil financial knowledge directly from those documents. As Japanese-language testbeds we adopt EDINET disclosures (Sugiura et al., 2026) and JP Fin Harness (Hirano, 2023)—EDINET measures performance on real-world financial tasks, while JP Fin Harness probes finance knowledge in Japanese—and additionally use GSM8K (Cobbe et al., 2021) to measure mathematical ability. Different benchmarks peak at different phases.

Contributions: **(1)** an empirical characterisation of phase-dependent capability trade-offs on EDINET and JP Fin Harness (Hirano, 2023), anchored to the un-fine-tuned baseline so each phase's contribution is decomposable; **(2)** a three-phase *staged continual adap-*

[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".

tation pipeline operating end-to-end on page images; **(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, 2025).

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); together with the vision encoder 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 Japanese government agencies—the Cabinet Office, FSA, and MOF—and publicly available math corpora (see Table 3, Appendix A). We chose the Cabinet Office, FSA, and MOF as the document sources because these agencies 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, 2025) as the teacher: **(i) an OCR dataset** that targets text transcription of each page, motivated by 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, motivated by LLaVAR-2 (Zhou et al., 2025); and **(iii) a finance-specific QA dataset** that probes finance knowledge tied to 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, 2025) 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 with XML-structured chain-of-thought supervision (full dataset details in Appendix A) and train the LLM (via LoRA-SFT) using Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025) 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; Zhu et al., 2021; Chen et al., 2022), 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. Hy-

[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

perparameters, 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) and after each of the three training phases. P0→P1 raises JP Fin Harness (45.0→54.2), GSM8K (58.5→64.6), and EDINET industry (4.7→8.5); EDINET earnings is essentially unchanged (0.524→0.532); EDINET fraud is the only benchmark that drops (0.516→0.472). From P1, GSM8K continues to rise to 73.2 (P2) before slipping to 71.9 (P3), and EDINET fraud recovers at P2 (0.534) and peaks at P3 (0.580). The remaining three benchmarks all decline past P1: JP Fin Harness (54.2→31.6→49.6), EDINET earnings (0.532→0.500→0.458), and EDINET industry (8.5→6.9→6.5). Relative to P0, P3 lies above P0 on four of five benchmarks (JP Fin Harness 49.6 vs. 45.0, GSM8K 71.9 vs. 58.5, EDINET fraud 0.580 vs. 0.516, EDINET industry 6.5 vs. 4.7) and below P0 only on EDINET earnings (0.458 vs. 0.524).

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). Because several of these numbers come from a different evaluation

Table 2. Evaluation against the un-fine-tuned baseline (P0) and after each training phase (P1–P3). † ROC-AUC; other metrics are Accuracy. Bold marks the best phase per row.

Benchmark	Phase 0 (P0)	Phase 1 (P1)	Phase 2 (P2)	Phase 3 (P3)
GSM8K	58.5	64.6	**73.2**	71.9
EDINET earnings†	0.524	**0.532**	0.500	0.458
EDINET fraud†	0.516	0.472	0.534	**0.580**
EDINET industry	4.7	**8.5**	6.9	6.5
JP Fin Harness	45.0	**54.2**	31.6	49.6

source (Sugiura et al., 2026), we treat the comparison as indicative context rather than a head-to-head result and defer the full tables to Appendix A.6. The one same-setup comparison is informative: our P3 checkpoint reaches 0.580 ROC-AUC on EDINET fraud, within 0.01 of the much larger Llama-3.3-70B (0.590), though it 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 is concentrated on `chabsa` (0.859→0.943), `cma_basics` (0.342→0.553), and `security_sales_1` (0.474→0.649). From P1, `chabsa` drops sharply at P2 (0.582) and nearly returns to P1 levels at P3 (0.936 vs. 0.943); `cma_basics`, `fp2`, and `security_sales_1` also drop at P2 but only partially recover at P3. `cpa_audit` is the only task whose best score is at P0 (0.276), with P3-1 (0.249) above P1 (0.226) but still below P0.

4. Discussion

Capability trajectories, not endpoints. No benchmark in our suite follows a flat trajectory: the best phase varies by benchmark (Table 2). JP Fin Harness, EDINET earnings, and EDINET industry peak at P1 (54.2, 0.532, 8.5); GSM8K peaks at P2 (73.2); EDINET fraud peaks at P3 (0.580). EDINET fraud is also the only benchmark that drops at P0→P1 (0.516→0.472) before recovering. Deployment checkpoint selection is therefore task-dependent: P1 for Japanese financial-domain priors, P2 for math, P3 for fraud detection. We treat staged continual adaptation as a *trajectory* anchored to P0 rather than a single endpoint.

Why P1→P2 erodes financial scores. JP Fin Harness drops sharply at P1→P2 (54.2→31.6), and EDINET earnings (0.532→0.500) and industry (8.5→6.9) decline as well; every JP Fin Harness subtask collapses, including evidence-grounded `chabsa` (0.943→0.582; Table 5). P2 trains exclusively on math problems with chain-of-thought supervision (Table 3), and we interpret these declines as catastrophic forgetting: heavy math-only supervision shifts the output distribution toward numerical chain-of-thought and overwrites the Japanese financial-domain priors acquired at P1. EDINET fraud is the exception (0.472→0.534): plausibly because the task relies heavily on numerical consistency-checking across accounting tables, for which math distillation provides directly relevant supervision, so the math gain appears to outweigh the lost domain priors.

Why P3 partially restores but does not match P1. P3 re-introduces financial supervision and lifts the finance benchmarks back: JP Fin Harness 31.6→49.6, `chabsa` 0.582→0.936, and EDINET fraud 0.534→0.580 (its peak). This shows that P3-style finance-specific adaptation is itself effective. However, JP Fin Harness, EDINET earnings, and EDINET industry all remain below their P1 peaks (49.6 vs. 54.2; 0.458 vs. 0.532; 6.5 vs. 8.5), and `chabsa` does not fully recover (0.943 vs. 0.936). We attribute the remaining gap primarily to the much smaller P3 supervision: P3-S1 (23,358) plus P3-S2 (32,484) totals only ∼56K examples (Table 3), an order of magnitude below the P1 corpus (1.08M for P1-S1; 365K for P1-S2) and roughly a third of the P2 corpus (∼150K problems). Scaling up P3 supervision is therefore the most direct lever to close the remaining gap.

What document-grounded VQA contributes. P1 is not a financial-task fine-tuning, but its training corpus is dominated by finance-related government PDFs, and the resulting P0→P1 jump on JP Fin Harness (45.0→54.2; Table 2) indicates that document-grounded VQA over these PDFs already populates non-trivial financial-domain priors. A complementary signal comes from P3: re-introducing visual domain QA from government PDFs at P3-S2, on top of the text-only financial fine-tuning at P3-S1, lifts the JP Fin Harness average from 0.4854 to 0.4959 (Table 5). Document-grounded VQA therefore has a measurable but modest effect on top of text-only adaptation; broader gains likely require a more deliberate VQA pipeline—e.g., a wider mix of question types beyond the extractive/reasoning split, more explicit table- and chart-grounded supervision, and harder distractors that force layout-aware comparison.

Limitations and scope. (i) All scores are single-seed runs, so we read inter-phase changes as trends rather than significance-tested effects. Our main conclusions rest on the large movements—the JP Fin Harness drop at P2 (54.2→31.6), the GSM8K gains (58.5→73.2), and the EDINET fraud recovery (0.472→0.580); smaller changes, such as the gradual EDINET earnings decline (0.532→0.500→0.458) and gaps below roughly 0.01 ROC-AUC, are directional only and may fall within seed variation, so confirming them would require multi-seed runs with confidence intervals. (ii) The results characterise a single model (an 8.4B SigLIP-v2 plus llm-jp backbone), a single language (Japanese), and a single domain (finance); we therefore read the phase-dependent trade-offs as a property of this specific staged pipeline, not as a general law of staged adaptation. (iii) Training is restricted to single-page inputs; multi-page training is a natural follow-up. (iv) A skip- or shorten-P2 ablation would test whether the fraud gain at P2 justifies the cost on industry and earnings. (v) P2 uses QLoRA while other stages use full-precision LoRA, so a small contribution from quantization and the P2→P3 adapter handoff cannot be excluded.

5. Conclusion

We presented an 8.4B-parameter multimodal foundation model for Japanese financial documents, trained via a synthetic-data pipeline over government-published PDFs and tracked from an un-fine-tuned baseline (P0) through three training phases. P0→P1 alone lifts JP Fin Harness from 45.0 to 54.2; across the staged trajectory the model reaches 0.580 ROC-AUC on EDINET fraud (at P3) and 73.2 on GSM8K (at P2). However, each benchmark peaks at a different phase: math distillation drives fraud gains but erodes EDINET industry, earnings, and taxonomy-heavy JP Fin Harness subtasks. Checkpoint selection is therefore task-dependent, and for the single model, language, and domain studied here, staged continual adaptation is best evaluated as a *trajectory* anchored to P0 rather than a single endpoint.

Acknowledgements

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

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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*, 2025.

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

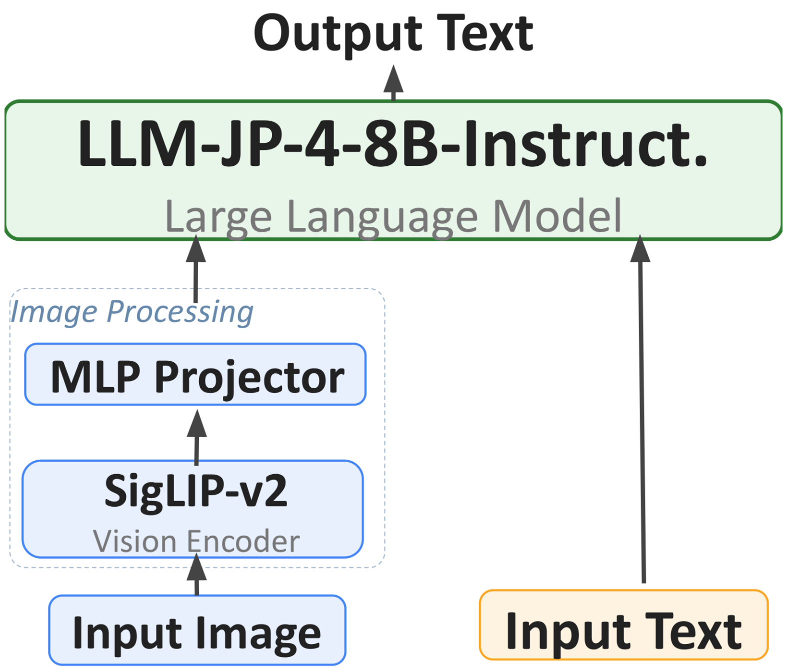


Figure 1. Architecture of the proposed model. The input image is connected to the language model through SigLIP-v2 and the MLP projector, and is processed jointly with the input text. By reading document pages directly as images, the model generates output text while preserving layout information such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

A.2. Staged Learning Pipeline Diagram

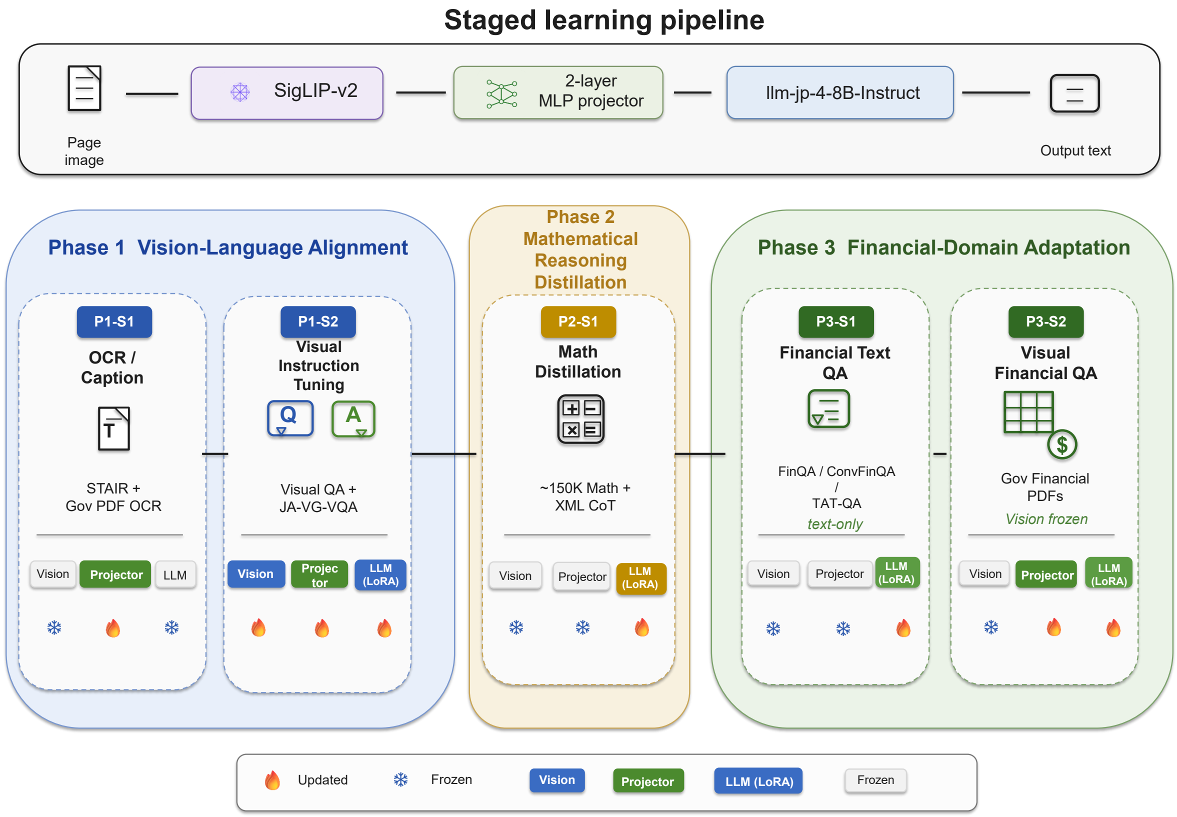


Figure 2. Overview of the three-phase staged continual adaptation pipeline. **Phase 1 (Vision-Language Alignment)**: P1-S1 trains only the projector on STAIR Captions and government-PDF OCR (vision encoder and LLM frozen); P1-S2 fine-tunes all three components (vision, projector, LLM via LoRA) on Visual QA and JA-VG-VQA. **Phase 2 (Mathematical Reasoning Distillation)**: P2-S1 updates the LLM via LoRA on approximately 50K 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)	540,000	
		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	23,358
		ConvFinQA train (Chen et al., 2022)	3,892	
		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. Bold indicates the best-performing phase for each task.

JP Fin task	P0	P1	P2	P3-1	P3-2
chabsa	0.8591	**0.9428**	0.5818	0.9357	0.9356
cma_basics	0.3421	**0.5526**	0.2105	0.3947	0.3947
cpa_audit	**0.2764**	0.2261	0.1683	0.2487	0.2362
fp2	0.2989	**0.3389**	0.1789	0.2863	0.2989
security_sales_1	0.4737	**0.6491**	0.4386	0.5614	0.6140
Average	0.4500	**0.5419**	0.3156	0.4854	0.4959

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.

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 result with the official 0-shot leaderboard (Hirano, 2023). On EDINET fraud, our 0.580 is within 0.01 of Llama-3.3-70B (0.590), evaluated under the same setup as ours, and is above the GPT-5 (0.560) and DeepSeek-R1 (0.540) numbers reported by EDINET-Bench (Sugiura et al., 2026) (the latter two comparisons are indicative since they draw on a different evaluation source). On EDINET earnings, our P1 peak (0.532) and our P3 checkpoint shown in Table 6 (0.458) both exceed Llama-3.3-70B (0.410) and the DeepSeek-R1 number reported by EDINET-Bench (0.430), while both trail GPT-5 (0.580). EDINET industry is a clear weakness: our 6.5 trails every external baseline (11.0–21.0). On JP Fin Harness, our P1 (54.2) exceeds Meta-Llama-3-8B-Instruct (44.7) and Qwen-14B-Chat (49.1) but trails Qwen2-72B-Instruct (67.7) and Claude 3.5 Sonnet (77.0); our P3 (49.6) is roughly at Qwen-14B-Chat parity (49.1). On GSM8K (Table 8), our P2 (73.2) exceeds Qwen-14B Chat (59.3) but remains 11–17 points below Llama 3 8B (84.5), Gemma 3 4B (89.2), and Qwen3 8B-base (89.84).

Table 6. EDINET comparison. Llama-3.3-70B (Meta, 2024) is evaluated under the same setup as ours; GPT-5 and DeepSeek-R1 numbers are taken 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	**0.580**	0.430	0.458
EDINET fraud†	**0.590**	0.560	0.540	0.580
EDINET industry	14.0	**21.0**	11.0	6.5

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) and Ours (P3) because JP Fin Harness peaks at P1. 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	54.2
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, 2025), Llama 3 8B (Grattafiori et al., 2024), and Gemma 3 4B (Gemma Team, 2025) results are taken from their respective technical reports. 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. Bold indicates the highest score.

Benchmark	Qwen-14B Chat	Qwen3 8B-base	Llama 3 8B	Gemma 3 4B	Ours (P0)	Ours (P1)	Ours (P2)	Ours (P3)
GSM8K	59.3	**89.84**	84.5	89.2	58.5	64.6	73.2	71.9

A.8. 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>
$\boxed{20}$
</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.9. Training Loss Curves

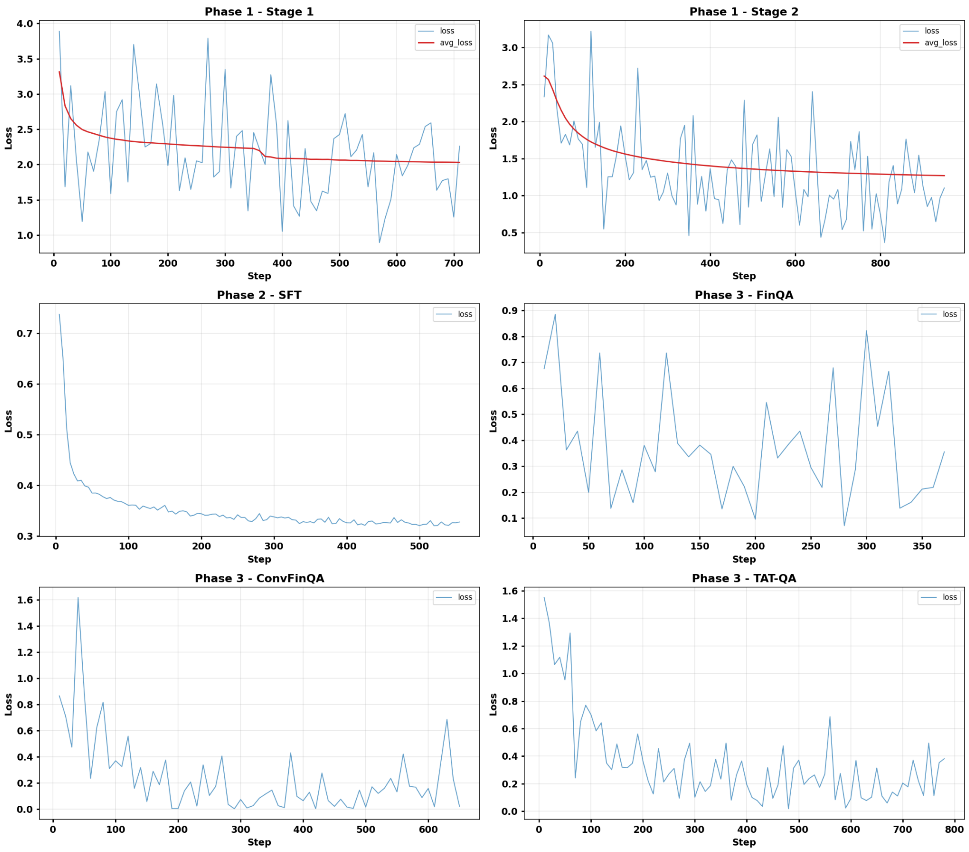


Figure 4. Loss curves during three-stage training. Phase 1 achieves vision–language alignment, Phase 2 develops reasoning, and Phase 3 performs financial adaptation in sequence.

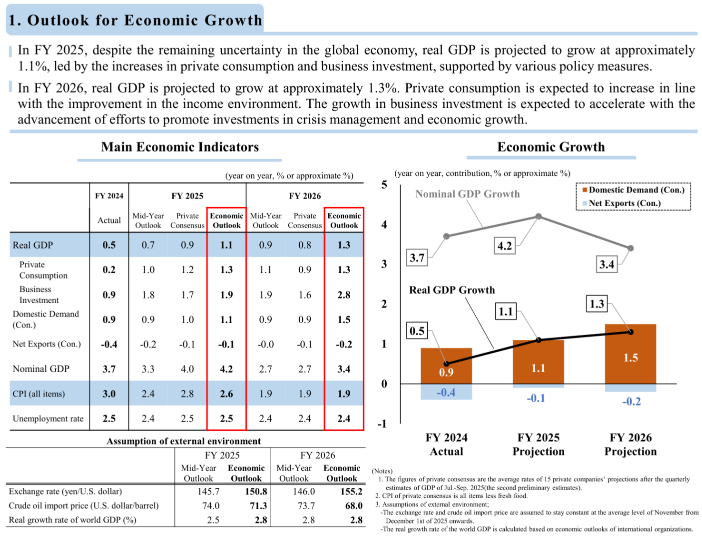
The loss curves in Figure 4 show stable decreases across Phase 1 (Stages 1 and 2), Phase 2 SFT, and Phase 3-1.

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 Actual	FY 2025 Mid-Year Outlook	FY 2025 Private Consensus	FY 2025 Economic Outlook	FY 2026 Mid-Year Outlook	FY 2026 Private Consensus	FY 2026 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 Mid-Year Outlook	FY 2025 Economic Outlook	FY 2026 Mid-Year Outlook	FY 2026 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

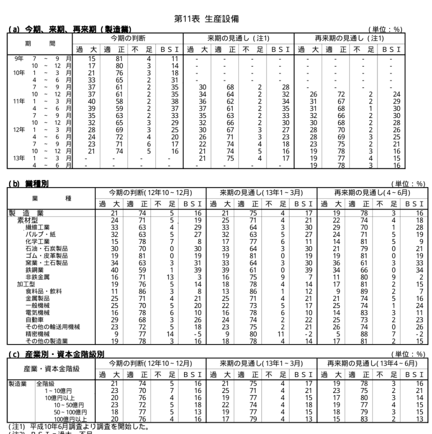
Figure 5. Source page.

```
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

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

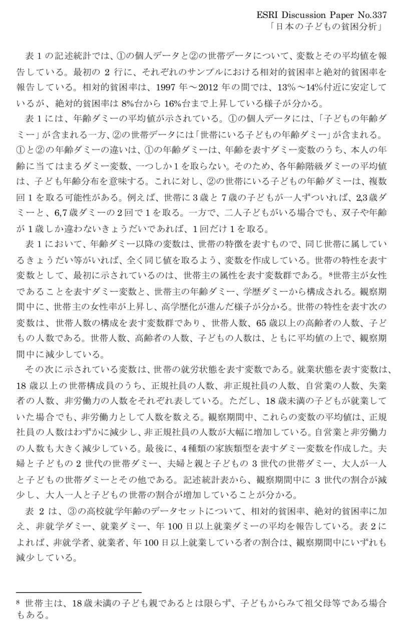
`Q:` From which cell was the "excessive" value for "Manufacturing" under "Current assessment (Oct–Dec 2012)" in Table (b) read?

`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。[8]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。

[8] 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.