# EuraGovExam: Can We Trust VLMs for Government Document Processing?

Anonymous Authors

January 27, 2026

## Abstract

As Vision-Language Models (VLMs) achieve expert-level performance on professional examinations, governments worldwide are actively exploring AI deployment for administrative document processing, citizen services, and examination grading. However, before deploying VLMs in high-stakes government systems, we must answer a critical question: *Can a model that excels on English benchmarks reliably process documents in Korean, Japanese, or Hindi?*

To address this question, we introduce **EuraGovExam**, a multilingual multimodal benchmark comprising 8,000+ civil service examination questions from five jurisdictions: Korea, Japan, Taiwan, India, and EU. Civil service exams are ideal for testing government AI readiness as they require jurisdiction-specific knowledge embedded in real scanned documents.

Our evaluation of 23 state-of-the-art VLMs reveals an alarming finding: **regional/language factors dominate task/subject factors by 3.9×** ($\eta^2 = 0.126$ vs 0.043, $p < 0.01$). GPT-4o achieves 63.7% on EU documents but plummets to 26.0% on Japanese documents—barely above random chance. This pattern is consistent across 82.6% of evaluated models.

These results challenge the assumption of "general intelligence" in VLMs and demonstrate that **region-specific validation is required** before government AI deployment.

# 1 Introduction

## 1.1 The Rise of VLMs and Government AI Adoption

Vision-Language Models (VLMs) have demonstrated remarkable progress in recent years. Performance on expert-level benchmarks has improved dramatically, from 56% to over 90% on MMMU within two years Yue et al. (2023). These models now achieve passing scores on professional examinations including medical licensing exams, bar examinations, and graduate-level entrance tests.

This rapid advancement has prompted governments worldwide to explore AI adoption for high-stakes administrative tasks. Applications under consideration include document processing, citizen service automation, examination grading, and regulatory compliance checking. The promise of increased efficiency and reduced costs makes AI deployment increasingly attractive to public sector organizations.

## 1.2 The Critical Question: Can We Trust Them?

Despite impressive benchmark performance, existing evaluations have significant limitations for assessing government AI readiness:

- **Language bias**: Most benchmarks are English-centric (MMMU, MathVista, DocVQA)
- **Clean inputs**: Synthetic or high-quality rendered images, not real scanned documents
- **Universal knowledge**: Mathematics, science, general knowledge—not jurisdiction-specific law or administration

**MMMU performance of 90% does not guarantee Korean government document processing capability.** Before deploying VLMs in high-stakes government systems, we must answer a critical question: *Can a model that excels on English benchmarks reliably process documents in Korean, Japanese, or Hindi?*

### 1.3 Our Finding: Regional Effects Dominate

We evaluated 23 state-of-the-art VLMs on EuraGovExam and discovered an alarming pattern: **regional performance variance is 3.9× larger than task performance variance.**

- Nation effect: $\eta^2 = 0.126$ (medium), $F(4, 110) = 3.95$, $p = 0.005$**
- Task effect: $\eta^2 = 0.043$ (small), $F(16, 368) = 1.05$, $p = 0.40$ (n.s.)
- 82.6% of models (19/23) show Nation > Task pattern (binomial $p = 0.0013$)

The most striking example: **GPT-4o achieves 63.7% on EU documents but drops to 26.0% on Japanese documents**—a 37.7 percentage point collapse to near-random levels. This is not an isolated case; the pattern persists across model families and sizes.

### 1.4 Why Civil Service Examinations?

Civil service examinations are uniquely suited for evaluating government AI readiness:

1. **Naturally occurring**: Official government-administered tests, not artificially constructed
2. **Ground truth available**: Official answer keys published by examination authorities
3. **Jurisdiction-specific**: Requires local law, administrative procedures, and regional knowledge
4. **Real documents**: Actual scanned examination papers with authentic visual characteristics

Unlike school examinations (EXAMS-V) which test transferable knowledge, civil service exams require understanding that cannot be obtained through translation alone.

### 1.5 Contributions

Our contributions are fourfold:

1. **Reliability Warning (Primary)**: We provide empirical evidence that VLM performance varies 3.9× more by region than by task, challenging assumptions of general intelligence.
2. **EuraGovExam Benchmark**: A multilingual multimodal benchmark with 8,000+ questions across 5 regions, 4 writing systems, and 17 subject domains.
3. **Diagnostic Analysis**: Taxonomy of failure modes revealing that 72% of failures are perception-related (OCR/script recognition), not reasoning failures.
4. **Practical Guidelines**: Region-specific recommendations for government AI deployment based on identified bottlenecks.

## 2 Related Work

### 2.1 Examination and Knowledge Benchmarks

**Text-only benchmarks** such as MMLU Hendrycks et al. (2020) evaluate broad knowledge but lack visual understanding requirements. **Multimodal benchmarks** including MMMU Yue et al. (2023) and MathVista Lu et al. (2023) test visual reasoning but use clean, rendered images and focus on universal knowledge domains.

**Multilingual exam benchmarks** like EXAMS-V Das et al. (2024) provide cross-lingual evaluation but test school-level examinations with transferable knowledge. In contrast, EuraGovExam tests jurisdiction-specific knowledge that cannot be acquired through translation.

### 2.2 Document Understanding Benchmarks

DocVQA Mathew et al. (2021), ChartQA Masry et al. (2022), and TextVQA Singh et al. (2019) evaluate document comprehension but are predominantly English-centric. OCRBench Liu et al. (2024b) specifically tests text recognition but does not combine OCR with domain reasoning requirements.

EuraGovExam uniquely combines: (1) real scanned documents with authentic visual characteristics, (2) multiple writing systems including complex scripts (Japanese, Korean, Devanagari), and (3) jurisdiction-specific domain knowledge requirements.

### 2.3 VLM Reliability Studies

Recent work has examined VLM reliability through perception analysis Chen et al. (2024), multilingual evaluation Liu et al. (2024a), and robustness testing. Our contribution differs in focus: we systematically demonstrate that regional factors dominate task factors in determining VLM performance, with direct implications for government AI deployment decisions.

## 3 EuraGovExam Dataset

### 3.1 Design Rationale

EuraGovExam is designed specifically for evaluating government AI readiness. Key design principles include:

- **Authenticity**: Real examination documents with original visual characteristics (scan noise, varied layouts)
- **Script diversity**: Four distinct writing systems to test multilingual capability
- **Jurisdiction specificity**: Content requiring local knowledge that cannot be obtained through translation

### 3.2 Data Sources

Table 1: EuraGovExam data sources by region

| Region | Source | Language | Script | Questions |
|---|---|---|---|---:|
| Korea | NPS () | Korean | Hangul | ∼2,000 |
| Japan | NPA () | Japanese | Kanji+Kana | ∼1,500 |
| Taiwan | MOEX () | Chinese | Traditional | ∼1,500 |
| India | UPSC | Hindi/English | Devanagari/Latin | ∼1,500 |
| EU | EPSO | EN/FR/DE | Latin | ∼1,500 |

### 3.3 Dataset Statistics

- **Total questions**: 8,000+
- **Regions**: 5 (Korea, Japan, Taiwan, India, EU)
- **Writing systems**: 4 (Hangul, Japanese mixed, Traditional Chinese, Latin/Devanagari)
- **Subject domains**: 17 (law, administration, economics, mathematics, sciences, etc.)
- **Time span**: 2015–2024
- **Format**: Multiple choice (4–5 options)

### 3.4 Quality Assurance

**Data quality**: Near-duplicate detection using perceptual hashing, PII removal verification, double-checked answer key alignment. Error rate <0.1%.

**Contamination analysis**: We analyze performance by examination year. Models show consistent patterns across temporal periods, including recent examinations (2023–2024) unlikely to be in training data.

## 4 Evaluation Protocol

### 4.1 Task Definition

We evaluate VLMs in an **image-only setting**: the model receives only the examination question image and must produce the correct answer choice (1–5). No external OCR tools or text extraction is permitted.

### 4.2 Evaluated Models

We evaluate 23 VLMs spanning closed and open-source models:

- **Closed (9)**: Gemini-2.5-pro, o3, o4-mini, GPT-4o, GPT-4.1, GPT-4.1-mini, Claude-Sonnet-4, Gemini-2.5-flash, Gemini-2.5-flash-lite
- **Open (14)**: Qwen2-VL (2B/7B/72B), InternVL2.5-38B, LLaVA variants, Ovis2 (8B/16B/32B), Phi-3.5-vision, Llama-3.2-11B-Vision

### 4.3 Metrics

- **Primary**: Accuracy (%)
- **Statistical**: 95% bootstrap confidence intervals, effect sizes ($\eta^2$, Cohen's $d$)
- **Significance**: ANOVA for factor analysis, paired t-tests with Bonferroni correction

## 5 Main Results

### 5.1 Overall Performance

Table 2 presents overall and regional performance for top models.

Table 2: Overall accuracy (%) on EuraGovExam. Regional scores show significant variance within each model.

| **Model** | **Overall** | **Taiwan** | **EU** | **Korea** | **India** | **Japan** | **Range** |
|---|---:|---:|---:|---:|---:|---:|---:|
| Gemini-2.5-pro | 87.0 | 95.5 | 88.1 | 91.1 | 69.2 | 87.6 | 26.3 |
| o3 | 84.3 | 93.7 | 84.5 | 90.1 | 68.6 | 82.4 | 25.1 |
| o4-mini | 79.4 | 92.3 | 77.0 | 82.5 | 63.4 | 82.5 | 28.9 |
| Gemini-2.5-flash | 68.3 | 92.7 | 83.3 | 67.7 | 62.3 | 51.5 | 41.2 |
| Claude-Sonnet-4 | 63.3 | 87.3 | 76.4 | 62.4 | 62.5 | 45.9 | 41.4 |
| GPT-4.1-mini | 56.3 | 79.0 | 63.6 | 59.9 | 46.3 | 43.8 | 35.2 |
| GPT-4.1 | 54.7 | 72.6 | 66.4 | 54.2 | 48.1 | 48.1 | 24.5 |
| Qwen2-VL-72B | 44.6 | 74.7 | 62.1 | 39.7 | 35.9 | 30.4 | 44.4 |
| **GPT-4o** | **42.0** | **66.7** | **63.7** | **33.2** | **41.0** | **26.0** | **40.7** |
| InternVL2.5-38B | 39.3 | 56.8 | 52.9 | 39.6 | 19.4 | 31.5 | 37.4 |

**Key observations**:

- Closed models significantly outperform open models (87.0% vs 44.6% for best in class)
- Regional variance within models often exceeds 40 percentage points
- Random baseline is 20–25%; some models approach random on difficult regions

### 5.2 Core Finding: Regional Effect Dominates Task Effect

We decompose performance variance using two-way ANOVA to test whether Nation or Task better explains performance differences.

Table 3: ANOVA results: Nation vs Task effects on VLM performance

| Factor | Variance | $\eta^2$ | F | p-value | Interpretation |
|---|---|---|---|---|---|
| Nation | 104.7 | 0.126 | 3.95 | 0.005** | Medium effect |
| Task | 27.0 | 0.043 | 1.05 | 0.40 | Small effect (n.s.) |
| **Variance Ratio: 104.7 / 27.0 = 3.9×** | | | | | |

**Core finding**: **Nation explains 3.9× more variance than Task.** The nation effect is statistically significant ($p < 0.01$) while the task effect is not ($p = 0.40$). This pattern holds for 82.6% of models (19/23), with binomial test $p = 0.0013$.

### 5.3 Regional Performance Hierarchy

Table 4: Regional difficulty ranking (averaged across all models)

| Rank | Region | Mean Acc | Std | Script | Characteristics |
|---|---|---|---|---|---|
| 1 (Hardest) | Japan | 32.5% | 23.4 | Kanji+Kana | Most complex script |
| 2 | India | 32.6% | 20.1 | Devanagari | Mixed scripts |
| 3 | Korea | 38.6% | 25.1 | Hangul | Syllabic script |
| 4 | EU | 49.9% | 23.3 | Latin | Familiar to models |
| 5 (Easiest) | Taiwan | 54.9% | 28.1 | Traditional Chinese | Best overall |

Performance gap between easiest and hardest regions: **22.4 percentage points**.

### 5.4 Evidence: Not Due to Exam Difficulty

A potential counterargument is that Japanese exams are inherently more difficult. We refute this by showing that **different models show different gaps on the same exams**:

- GPT-4o: Taiwan 66.7% → Japan 26.0% (**40.7pp drop**)
- Gemini-2.5-pro: Taiwan 95.5% → Japan 87.6% (**7.9pp drop**)

If Japanese exams were inherently harder, both models should show similar proportional drops. The 5× difference in drop magnitude demonstrates that the gap reflects **model capability differences**, not exam difficulty.

## 6 Diagnostic Analysis

### 6.1 Failure Taxonomy

We manually analyzed 176 failure cases to categorize error types.

**Key insight**: **72.2% of failures are perception-related** (script, OCR, symbols), while only **11.4% are pure reasoning failures**. VLMs fail primarily because they "cannot read," not because they "cannot think."

### 6.2 Regional Bottleneck Patterns

Different regions exhibit distinct failure patterns:

- **Japan & Korea**: Perception bottleneck (script recognition >40% of failures)

Table 5: Primary failure categories across all regions

| **Category** | **Count** | **%** | **Type** |
|---|---:|---:|---|
| Vertical/non-Latin script | 71 | 40.3% | Perception |
| OCR/text recognition | 31 | 17.6% | Perception |
| Math symbol interpretation | 25 | 14.2% | Perception |
| Pure reasoning/knowledge | 20 | 11.4% | Reasoning |
| Diagram understanding | 12 | 6.8% | Perception |
| Table structure | 11 | 6.3% | Perception |
| Other | 6 | 3.4% | Mixed |

- **India**: Perception bottleneck (OCR + script issues account for 97% of failures)
- **EU**: Reasoning bottleneck (48% of failures are knowledge/reasoning-related)
- **Taiwan**: Minimal failures; when present, mostly OCR-related

This suggests that Asian language performance can be significantly improved through better OCR capabilities, while EU performance requires enhanced domain knowledge.

### 6.3 Model-Specific Insights

**Gemini vs GPT comparison** reveals stark differences in Asian language handling:

Table 6: Gemini vs GPT performance by region

| **Region** | **Gemini-2.5-pro** | **GPT-4o** | **Gap** |
|---|---:|---:|---:|
| Japan | 87.6% | 26.0% | +**61.6pp** |
| Korea | 91.1% | 33.2% | +**57.9pp** |
| India | 69.2% | 41.0% | +28.2pp |
| Taiwan | 95.5% | 66.7% | +28.8pp |
| EU | 88.1% | 63.7% | +24.4pp |

Gemini shows consistent performance across regions (69–95%), while GPT-4o exhibits dramatic variance (26–67%). The 61.6pp gap on Japanese documents suggests fundamental differences in Asian language processing capability.

## 7 Discussion

### 7.1 Implications for Government AI Deployment

Our findings have direct implications for government AI adoption:

1. **Region-specific validation is mandatory**: English benchmark performance does not predict performance on Asian language documents. Governments must evaluate on target-region data before deployment.
2. **Model selection matters significantly**: A 3× performance difference is possible between model families on the same task (Gemini 87.6% vs GPT-4o 26.0% on Japanese).
3. **Human-in-the-loop required for high-stakes decisions**: Given performance variability, fully automated decision-making is not recommended for legal, administrative, or citizen-impacting tasks.

### 7.2 Why Do Regional Gaps Exist?

We hypothesize four contributing factors:

1. **Training data imbalance**: English data dominates VLM training; Asian language document data is scarce

2. **OCR quality gap**: Built-in OCR is optimized for Latin scripts
3. **Jurisdiction knowledge gap**: US/EU law knowledge exceeds Asian law knowledge
4. **Benchmark bias**: Existing benchmarks are English-centric, reducing incentive for multilingual improvement

### 7.3 Limitations

- **Geographic coverage**: 5 regions; Southeast Asia, Middle East, Africa not included
- **Contamination**: Some older exams may be in training data; however, patterns persist on recent (2023–2024) exams
- **Causality**: Observational study; causal mechanisms require further investigation
- **Task scope**: Multiple choice only; free-form response evaluation is future work

## 8 Conclusion

We introduced EuraGovExam, a multilingual multimodal benchmark for evaluating VLM reliability on government document processing. Our evaluation of 23 VLMs reveals that **regional factors dominate task factors by 3.9**× in determining performance. GPT-4o drops from 63.7% (EU) to 26.0% (Japan)—near random chance—demonstrating that high English benchmark scores do not guarantee multilingual reliability.

Our diagnostic analysis shows that 72% of failures are perception-related (OCR, script recognition), suggesting that improved multilingual OCR could substantially enhance Asian language performance.

**Key message**: Governments seeking to deploy VLMs must conduct region-specific validation. The assumption of "general intelligence" is not supported by empirical evidence. EuraGovExam provides a standardized tool for this critical evaluation.

#### Acknowledgments

[To be added]

## References

W. Chen, X. Zhu, M. Shi, Y. Zhang, W. Chen, Z. Ding, Y. Li, and X. Wan. Can VLMs see beyond visual perception? A case study on complex reasoning. *arXiv preprint arXiv:2402.12982*, 2024.

R. J. Das, O.-M. Camburu, R. Aralikatte, E. Ivanova, O. Kursun, J. Barrow, R. Vedantam, K.-W. Lee, and N. Lee. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Y. Liu, Y. Dou, Y. Wang, K. Zhang, Y. Wang, and W. Fan. Multilingual evaluation of vision-language models. *arXiv preprint*, 2024a.

Y. Liu, Z. Li, H. Bai, Y. Liu, J. Liu, C. Zhang, Z. Zheng, and B. Lin. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2024b.

P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

M. Mathew, D. Karatzas, and C. Jawahar. DocVQA: A dataset for VQA on document images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021.

A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. TextVQA: Towards reading text in images to answer questions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *arXiv preprint arXiv:2311.16502*, 2023.