Structured Entity Extraction from Hawaiian Television Chyrons Using Vision-Language Models

Kelley Lynch[1], **Owen King**[2], **Kyeongmin Rim**[1], **Gabrielle Keen**[2], **Yangyang Chen**[1], **James Pustejovsky**[1]
[1]Department of Computer Science, Brandeis University
[2]GBH Archives
{kmlynch, krim, yangyangchen, jamesp}@brandeis.edu
{owen_king, gabrielle_keen}@wgbh.org

Abstract

Hawaiian (ʻŌlelo Hawaiʻi) is an endangered Polynesian language whose broadcast archives represent a critical yet underutilized resource for language documentation. We present the first evaluation of vision-language models (VLMs) for structured entity extraction from television chyrons, investigating the performance gap between Hawaiian-language content and mainland U.S. comparisons. Using our new HiChy dataset of 3,925 manually annotated images, we demonstrate that Hawaiian content remains significantly more challenging for current VLMs: for the best-performing model (Qwen2.5-VL-7B), character error rates roughly double from 0.064 on mainland data to 0.130 on Hawaiian content. We extend the task to key information extraction (KIE), finding that while models can perform structured parsing, they struggle specifically with names of Hawaiian linguistic origin, a difficulty that persists even when controlling for geographic source. Across five evaluated models spanning local quantized inference and commercial APIs, we find that OCR accuracy and structured extraction capability do not necessarily correlate: the best OCR model (Gemini 3 Flash) underperforms locally-deployed alternatives on KIE, while even a 2.2B-parameter model (SmolVLM2) achieves functional extraction. Our results provide a baseline for AI-assisted archival processing of underrepresented language media and highlight the need for models that better account for the orthographic and cultural specificities of Hawaiian.

Keywords: key information extraction, vision-language models, chyron OCR, Hawaiian language, broadcast archives, low-resource languages

1. Introduction

Lower-third graphics, commonly known as chyrons, are a primary means of identifying individuals in broadcast television. They typically display a person's name along with titles, affiliations, or other descriptive attributes. For archival institutions managing large collections of historical broadcast video, the text in these graphics represents a rich source of structured metadata (names, roles, and organizational affiliations) that could support cataloging, search, and reconciliation with name authority files.

This work focuses on content from PBS Hawaiʻi, a public television station whose archives contain decades of programming featuring Hawaiian-language names and cultural content. Hawaiian (*ʻŌlelo Hawaiʻi*) is an endangered Polynesian language that has been the subject of intensive revitalization efforts since the founding of the ʻAha Pūnana Leo Hawaiian-medium preschools in 1984 (Wilson and Kamanā, 2001). While the speaker community has grown through immersion education, Hawai-

(a) Hawaiian: name with title.

(b) Hawaiian: attributes and background text.

(c) Comparison: mainland U.S. broadcast.

Figure 1: Example chyron images from the HiChy dataset showing Hawaiian (a, b) and mainland U.S. comparison (c) subsets. The KIE task requires extracting the person's name, normalizing it, and identifying additional attributes.

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 160–168
11–12 May 2026.

ian remains classified as endangered (Eberhard et al., 2025) and computational resources for the language are extremely limited. Prior NLP work on Hawaiian includes character-level models for recovering missing diacritical marks in historical texts (Shillingford and Parker Jones, 2018), and broader surveys have highlighted the scarcity of OCR evaluation data for Indigenous and Pacific Island languages (Agarwal and Anastasopoulos, 2024). Broadcast archives represent an underexploited resource for both archival access and language documentation, making structured extraction from these materials a task of particular relevance to the language technology community.

Prior work has evaluated vision-language models for the task of *transcribing* chyron text (OCR) within AI-assisted archival processing platforms (Rim et al., 2025), demonstrating that modern VLMs can capture the semantic content of chyrons despite challenges with exact case preservation and formatting. However, raw transcription alone is insufficient for many archival applications. The more practically valuable task is *structured entity extraction*: parsing chyron text into discrete, semantically meaningful fields that can be directly integrated into metadata records.

In this paper, we evaluate VLMs on key information extraction (KIE) from chyron images. Given a chyron image, the model must produce a structured output containing: (1) the person's name exactly as displayed (*name-as-written*), (2) the name in a normalized "Lastname, Firstname" format (*name-normalized*), and (3) a list of additional attributes such as titles, roles, or affiliations. This task combines visual text recognition with lightweight reasoning, as the model must not only read the text but also identify which portion is a name, reformat it, and categorize remaining text as attributes.

We evaluate five VLMs on the HiChy dataset, which comprises 3,925 annotated chyron images from Hawai'i public television and mainland U.S. broadcasts. This dataset includes names from Hawaiian, Japanese, and English linguistic traditions. It specifically features the vowel-dense phonotactic patterns and open-syllable structures characteristic of Polynesian languages. These present significant challenges for models trained predominantly on English text because they represent out-of-distribution character sequences that diverge from Western-centric training data.

Our contributions are: (1) the first evaluation of VLMs for structured entity extraction from broadcast chyrons; (2) a comparison of five models spanning local quantized inference and commercial APIs, revealing that local mid-scale models can outperform API-based alternatives for structured extraction; (3) an ablation comparing few-shot and zero-shot prompting, showing that examples primarily help with name normalization conventions; and (4) analysis of how name origin and orthographic diversity affect extraction difficulty.

2. Related Work

2.1. Key Information Extraction

Key information extraction (KIE) from document images is a well-studied task in the document AI literature. Standard benchmarks include SROIE (Huang et al., 2019) for receipt parsing (4 fields: company, date, address, total), FUNSD (Jaume et al., 2019) for form understanding, and CORD (Park et al., 2019) for receipt parsing with 30+ fields. These benchmarks evaluate models on extracting structured information from visually rich documents, typically using entity-level F1 or field-level exact match as metrics.

More recently, ANLS (Average Normalized Levenshtein Similarity) has emerged as the preferred metric for evaluating text-generating models on document understanding tasks (Biten et al., 2019). ANLS applies a soft matching threshold ($\tau = 0.5$), giving partial credit for near-correct answers while assigning zero to predictions below the similarity threshold. ANLS* extends this to structured outputs by aligning predicted and gold fields and averaging ANLS across them (Peer et al., 2024).

While KIE has been extensively studied for receipts, forms, and business documents, broadcast television graphics have received little attention. The VKIE benchmark (An et al., 2024) addresses key information extraction from video text but focuses on contemporary digital content. Our work extends KIE evaluation to historical broadcast graphics, a domain characterized by analog degradation, varied typography, and culturally specific content.

2.2. Vision-Language Models for OCR

Vision-language models have demonstrated strong performance on OCR tasks, often surpassing traditional OCR pipelines on complex visual text. OCRBench v2 (Fu et al., 2025) provides comprehensive evaluation across 31 scenarios, finding that most current models still struggle with fine-grained perception and complex element parsing. The NewsVideoQA dataset (Jahagirdar et al., 2023) demonstrates the value of combining visual and textual cues for understanding news video content.

For chyron-specific OCR, a companion paper evaluating LLaVA and SmolVLM2 on Hawaiian broadcast content finds that VLMs significantly outperform traditional OCR tools (docTR, Tesseract) but struggle with case preservation and formatting. Our work builds on these OCR baselines by extend-

ing evaluation to the more demanding KIE task with new models and standard document AI metrics.

2.3. NER for Low-Resource Languages

Named entity recognition for low-resource languages remains challenging. Multilingual NER benchmarks such as Universal NER (Mayhew et al., 2024), which covers 37 languages, and MasakhaNER 2.0 (Adelani et al., 2022), which targets 20 African languages, have expanded coverage beyond high-resource settings but still exclude Hawaiian and most Pacific Island languages. OCR for low-resource languages faces similar gaps: Agarwal and Anastasopoulos (2024) survey the field and find that most evaluation efforts focus on South and Southeast Asian scripts, with no coverage of Polynesian languages. The HiChy dataset represents the first structured entity extraction resource for Hawaiian-language broadcast content, contributing to the broader effort of developing NLP tools for underrepresented languages.

3. Dataset

3.1. HiChy: Hawaiian Chyron Dataset

The HiChy dataset comprises manually annotated chyron images from two sources:

- **Hawaiian (HI)**: 1,825 chyrons from 158 PBS Hawai'i broadcast programs spanning 1975–2008, covering news and talk show formats.
- **Comparison (comps)**: 2,100 chyrons from continental U.S. public television broadcasts across 131 programs, providing a baseline for cross-dataset comparison. This subset was curated using the same extraction and verification methodology as the Hawaiian data.

The comparison programs were selected by an archivist to cover similar time periods (1980s through 2000s) and formats as the Hawaiian data, yielding comparable distributions of chyron styles and text complexity between the two subsets. Candidate chyron frames were identified using a ConvNeXt-based classifier (Liu et al., 2022) and manually verified to remove false positives. Each chyron image is annotated with: (1) a raw transcription of the on-screen text, (2) the person's name as written, (3) a normalized name in "Lastname, Firstname" format, and (4) a list of additional attributes (titles, roles, affiliations).

3.2. Dataset Characteristics

The Hawaiian subset presents distinct linguistic challenges not found in the comparison data. The names in this subset feature the vowel-dense phonotactic patterns and open-syllable structures characteristic of Polynesian languages. These represent out-of-distribution character sequences that diverge significantly from the English-centric patterns dominant in VLM training sets. Additionally, historical broadcasts feature analog artifacts, varied typography, and culturally specific visual elements such as leis and floral patterns that can occlude or overlap with chyron text (Figure 2).

Statistic	HI	Comps
Chyrons	1,825	2,100
Avg. chars/chyron	38.1	35.0
Avg. words/chyron	5.5	5.1
Avg. lines/chyron	3.1	2.9
% with attributes	79.3%	84.5%
% with diacritics	0.8%	0.0%
Name origin		
Hawaiian (H)	225	0
Japanese (J)	195	5
Both (B)	34	0
Neither (N)	1,371	2,095

Table 1: Dataset statistics for the Hawaiian and comparison subsets.

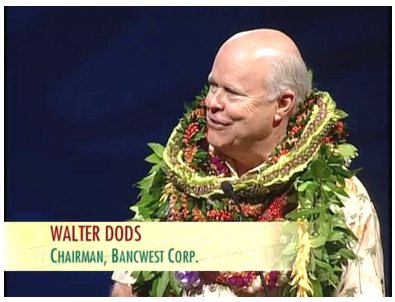
To isolate specific factors that may affect extraction difficulty, each name was labeled by linguistic origin: Hawaiian (H, n=225), Japanese (J, n=200), both (B, n=34), or neither (N, n=3,466), using an AI-assisted, human-verified procedure.

4. Method

4.1. Models

We evaluate five vision-language models spanning local and API-based deployment:

- **Qwen2.5-VL-7B** (Qwen Team, 2025): A 7-billion parameter VLM from the Qwen2.5 family, with strong multilingual and OCR capabilities.
- **Qwen2.5-VL-32B** (Qwen Team, 2025): The 32-billion parameter variant, providing a scaling comparison within the same architecture.
- **Gemini 3 Flash** (Google DeepMind, 2025): A commercial VLM accessed via the Google Gemini API, included to compare local quantized models against API-based alternatives.
- **Nanonets-OCR2** (3B): An OCR-specialized model fine-tuned from Qwen2.5-VL-3B (Mandal et al., 2025).
- **SmolVLM2** (2.2B): A compact, efficient VLM designed for resource-constrained deployment (Marafioti et al., 2025).



(a) Leis overlapping the chyron region.

(b) Hawaiian shirt pattern behind text.

(c) Foliage background with Hawaiian name.

Figure 2: Examples of challenging Hawaiian broadcast chyrons. Cultural elements such as leis, floral shirt patterns, and foliage can occlude or overlap with chyron text, compounding the difficulty of text extraction.

The Qwen models and Nanonets-OCR2 are served via Ollama using quantized (GGUF) weights on a single NVIDIA RTX A6000 GPU (48GB VRAM), reflecting the computational constraints typical of archival institutions. SmolVLM2 is run via HuggingFace Transformers with full-precision weights on the same hardware. Gemini 3 Flash is accessed via the Google API with identical prompts.

4.2. Task Definitions

We evaluate two tasks:

OCR (Task 1): Raw transcription of chyron text. The model receives a chyron image and must output the text exactly as displayed, with line breaks represented as `\n`.

KIE (Task 2): Structured entity extraction. The model receives a chyron image and must output a JSON object with three fields: `name-as-written`, `name-normalized`, and `attributes`.

4.3. Prompting Strategy

For OCR, we use a system prompt establishing the model as a precise text transcription system and a user prompt directing attention to the lower-third region.

For KIE, we use a system prompt establishing the model as an expert OCR and information extraction system, and a user prompt specifying the JSON schema with field definitions. We evaluate two variants:

- **Few-shot** (3 examples): The user prompt includes three demonstrations showing the expected input-output mapping, including one example with a Hawaiian name containing an *'okina*.
- **Zero-shot**: Identical instructions but without examples, ending with "Output only the JSON object, nothing else."

Full prompts are provided in Appendix A.

4.4. Evaluation Metrics

Following standard practice in the document AI literature, we adopt the following metrics:

OCR metrics: Character Error Rate (CER), Word Error Rate (WER), Exact Match (EM), and Average Normalized Levenshtein Similarity (ANLS, $\tau = 0.5$). All metrics are computed in both case-sensitive and case-insensitive variants.

KIE metrics: Per-field ANLS, EM, and CER for each of the three structured fields. We report ANLS* (field-aligned average ANLS) as our primary aggregate metric (Peer et al., 2024). Case-insensitive variants are also computed.

Cross-cutting analyses: We report metrics broken down by dataset (Hawaiian vs. comparison) and by presence of diacritical marks in the gold text.

5. Results

5.1. OCR Performance

Table 2 presents OCR results across all five models.

Four models produce meaningful OCR (Table 2). Gemini 3 Flash leads on all metrics, followed by Qwen 7B and SmolVLM2. The 32B Qwen model underperforms its 7B counterpart due to word-repetition artifacts (e.g., "LYNETTE LO TOM TOM"). Nanonets-OCR2 produces near-total failures due to hallucinated content and degenerate output loops.

Case-insensitive evaluation reveals that much of the apparent error is case mismatch: Gemini's

Model	CER↓	WER↓	EM↑	ANLS↑
Gemini Flash	**0.076**	**0.080**	**0.157**	**0.910**
Qwen 7B	0.095	0.159	0.148	0.895
SmolVLM2	0.116	0.185	0.135	0.867
Qwen 32B	0.193	0.244	0.085	0.825
Nanonets	0.803	0.842	0.029	0.108

Table 2: OCR results (case-sensitive). CER and WER are capped at 1.0 per instance before averaging.

Model	NAW	NN	Attr.	ANLS*
Qwen 32B	0.848	0.942	**0.828**	**0.873**
Qwen 7B	0.822	**0.967**	0.822	0.870
Gemini Flash	0.747	0.747	0.775	0.756
SmolVLM2	**0.861**	0.672	0.701	0.745
Nanonets	0.000	0.000	0.179	0.060

Table 3: KIE results (case-sensitive). Per-field scores are ANLS for name-as-written (NAW), name-normalized (NN), and attributes (Attr.); the aggregate is ANLS*.

ANLS jumps from 0.910 to 0.993, indicating near-perfect text reading but systematic title-case normalization regardless of on-screen capitalization. All capable models show similar improvements under case-insensitive evaluation, confirming that case mismatch between ALL CAPS chyrons and gold annotations is a consistent error source.

5.2. KIE Performance

Table 3 presents KIE results with few-shot prompting.

For KIE, four models produce functional structured outputs. The Qwen models achieve the highest ANLS* scores (32B: 0.873, 7B: 0.870), with complementary strengths: the 7B model excels at name normalization (0.967 vs. 0.942) while the 32B model is slightly better at name-as-written extraction (0.848 vs. 0.822). Both models achieve strong attribute extraction ($\sim$0.82 ANLS).

Gemini 3 Flash and SmolVLM2 2.2B achieve similar aggregate scores (ANLS* 0.756 and 0.745 respectively) but with very different error profiles. SmolVLM2 achieves the best name-as-written extraction of any model (0.861), demonstrating strong visual text reading, but struggles with name normalization (0.672), suggesting that the "Lastname, Firstname" reformatting convention is difficult for a 2.2B model to learn from few-shot examples alone. Gemini 3 Flash shows uniformly moderate performance across all three fields ($\sim$0.75), with its case normalization behavior depressing the name-as-written score.

Case-insensitive evaluation improves scores across models: the Qwen 7B ANLS* rises from 0.870 to 0.915, SmolVLM2 from 0.745 to 0.877, and Gemini 3 Flash from 0.756 to 0.809, confirming that case mismatch is a consistent error source.

Model	Prompt	NAW	NN	Attr.	ANLS*
Qwen 7B	few-shot	0.822	**0.967**	0.822	0.870
	zero-shot	0.752	0.935	0.799	0.829
Qwen 32B	few-shot	**0.848**	0.942	**0.828**	**0.873**
	zero-shot	0.840	0.861	0.807	0.836

Table 4: Few-shot vs. zero-shot KIE ablation (case-sensitive).

Model	Split	CER↓	ANLS↑	ANLS*↑
Gemini Flash	comps	**0.041**	**0.960**	0.883
	hi	**0.116**	**0.852**	0.611
Qwen 7B	comps	0.064	0.936	0.897
	hi	0.130	0.847	**0.840**
Qwen 32B	comps	0.151	0.878	**0.908**
	hi	0.242	0.764	0.832
SmolVLM2	comps	0.095	0.897	0.805
	hi	0.141	0.833	0.676

Table 5: Performance by dataset. CER and ANLS are OCR metrics; ANLS* is the KIE aggregate.

5.3. Few-Shot vs. Zero-Shot KIE

Table 4 compares few-shot (3 examples) and zero-shot KIE prompting for the Qwen models.

Few-shot examples improve ANLS* by 4.1 points for the 7B model and 3.7 points for the 32B model. The largest gains are in name normalization: the examples teach the "Lastname, Firstname" convention, yielding +3.2 and +8.1 point improvements for 7B and 32B respectively. Name-as-written and attribute extraction show smaller but consistent improvements, suggesting that examples also help with output formatting and field boundary identification.

5.4. Cross-Dataset Analysis

Table 5 confirms that Hawaiian content is consistently harder across all models and both tasks. The OCR gap is substantial: CER roughly doubles on Hawaiian content across most models (e.g., Gemini: 0.041 $\rightarrow$ 0.116; Qwen 7B: 0.064 $\rightarrow$ 0.130). The KIE gap is even more pronounced for Gemini (ANLS* drops from 0.883 to 0.611), while the Qwen models degrade more gracefully.

		Control	Name Origin					
Model	**Metric**	**Neither**	**Haw.**	**Δ**	**Jpn.**	**Δ**	**Both**	**Δ**
Qwen 7B	ANLS*	0.873	0.829	−0.044	0.873	0.000	0.893	+0.020
	CER	0.091	0.147	+0.056	0.107	+0.016	0.079	−0.012
Qwen 32B	ANLS*	0.877	0.814	−0.063	0.871	−0.006	0.844	−0.033
	CER	0.191	0.239	+0.048	0.193	+0.002	0.137	−0.054
Gemini Flash	ANLS*	0.789	0.475	−0.314	0.599	−0.190	0.275	−0.514
	CER	0.073	0.132	+0.059	0.074	+0.001	0.029	−0.044
SmolVLM2	ANLS*	0.753	0.619	−0.134	0.746	−0.007	0.700	−0.053
	CER	0.113	0.173	+0.060	0.119	+0.006	0.060	−0.053

Table 6: Performance by name origin relative to the Neither (control) baseline. Δ shows the difference from the control. For ANLS* (KIE aggregate), negative Δ indicates worse performance; for CER (OCR error), positive Δ indicates worse performance.

5.5. Impact of Name Origin

To isolate whether the difficulty is linguistic rather than geographic, Table 6 breaks down performance by name origin. Every model shows a consistent Hawaiian penalty (Δ column). Gemini Flash is most affected (ANLS* drops 0.314 points); the Qwen models are more robust (ANLS* gaps of 0.044–0.063). Japanese names fall between Hawaiian and neither-origin names across all models.

5.6. Orthographic Discrepancy

A notable characteristic of the HiChy dataset is the near-total absence of diacritical marks. While many names in the dataset are traditionally written with the *ʻokina* and *kahakō*, these characters appear in only 0.4% of the gold annotations. This omission likely reflects historical technical limitations in broadcast graphics generation. Because our gold standard is a transcription of the text as it appears on screen, a model that successfully restores these marks would technically be penalized for an incorrect match. This underscores a fundamental tension in archival processing: a model may produce a linguistically "correct" spelling that nonetheless fails to match the visual evidence of the historical record. Future work should explore diacritical restoration as a specialized post-processing step rather than an expectation of the raw extraction task.

6. Discussion

6.1. OCR vs. KIE Tradeoffs

OCR capability and KIE capability diverge across models. Gemini 3 Flash achieves the best raw OCR (case-insensitive ANLS=0.993) but the weakest KIE among capable models (ANLS*=0.756), while Qwen2.5-VL-7B achieves the best KIE (ANLS*=0.870) despite slightly worse OCR. Notably, SmolVLM2 at just 2.2B parameters produces competitive OCR (ANLS=0.867) and functional KIE (ANLS*=0.745), including the best name-as-written extraction of any model (0.861), though its weaker name normalization (0.672 vs. 0.967 for 7B) suggests that the "Lastname, Firstname" reformatting convention is harder to learn from few-shot examples at smaller scales. However, since SmolVLM2 was run with full-precision weights while the Qwen models used 4-bit quantization, and we evaluated only direct prompting without chain-of-thought or model-specific optimization, we cannot draw firm conclusions about parameter-count requirements. These results suggest that structured extraction depends more on following formatting conventions than on raw OCR accuracy. The name origin analysis (Table 6) further reveals that the Hawaiian performance gap is linguistic, not geographic: Hawaiian names are harder than names of neither origin *even within the Hawaiian broadcast subset*, confirming that the difficulty stems from unfamiliarity with Hawaiian orthographic patterns.

6.2. Error Analysis

Table 7 presents representative examples of the error types observed across models. Error analysis reveals distinct failure modes for each model family.

Qwen models. For the 7B model, *field boundary confusion* (18.5% of examples) is the most frequent error: the model places all chyron text, including attributes, into the name-as-written field, or splits compound attributes at internal punctuation. The 32B model's dominant error is *repetition artifacts* (37.4%), producing outputs like "LYNETTE LO TOM TOM" that cascade into normalization errors. These repetition artifacts may be exacerbated by 4-bit quantization, which can amplify token-level sampling instabilities in larger models. Both models struggle with unconventional layouts where the displayed name is already in "Lastname, Firstname" format, causing field mismapping.

Gemini 3 Flash. Despite strong OCR, Gemini exhibits three KIE-specific issues: (1) *case preservation*: Gemini faithfully reads ALL CAPS text, but gold annotations normalize to title case, producing ANLS=0.0 on correctly read text (234 examples); (2) *JSON parsing failures*: 19% of KIE predictions cannot be parsed, often due to multiline JSON in markdown code blocks; and (3) *background text inclusion*: Gemini reads program or station branding (such as channel logos) alongside chyron text. We note that case-insensitive evaluation substantially improves Gemini's scores (OCR ANLS: 0.910→0.993; KIE ANLS*: 0.756→0.809), confirming that much of the gap reflects formatting conventions rather than recognition errors. We retain case-sensitive evaluation as primary because case preservation is important for archival metadata fidelity.

SmolVLM2 2.2B. Despite strong name-as-written extraction (0.861 ANLS), SmolVLM2 struggles with the normalization step, frequently producing names in "Firstname Lastname" order rather than the required "Lastname, Firstname" format, or including titles that should be stripped. This suggests that the 2.2B model can read text accurately but has limited capacity for learning formatting conventions from few-shot examples.

Nanonets-OCR2. This model generates degenerate loops and hallucinated text bearing no relation to the input image, producing non-functional output for both tasks. Since the Qwen models use the same Ollama/GGUF quantization pipeline without issue, the failure is likely attributable to domain mismatch: Nanonets-OCR2 was fine-tuned on clean digital documents (receipts, forms), making it brittle when applied to analog broadcast graphics with varied typography and visual noise.

Linguistic errors. Models occasionally produce English-like phonetic approximations of Hawaiian names, and none reliably produce the *ʻokina* or *kahakō* when present in gold annotations. Since broadcast graphics themselves routinely omit these marks, the models' inability to restore them is unsurprising but underscores the gap between displayed text and correct Hawaiian orthography.

6.3. Implications for Archival Workflows

The ANLS* scores of ∼0.87 for the Qwen models indicate that structured extraction is viable for archival metadata generation in human-in-the-loop workflows. The high name normalization scores (NN ANLS $>$ 0.94) are especially promising for integration with name authority files. Notably, the best results come from locally-deployed models, relevant for institutions with privacy concerns or limited API budgets. Even SmolVLM2 (2.2B, ∼4.5GB) could serve institutions with limited GPU resources.

6.4. Limitations

Our evaluation is limited to five models with varying inference configurations: the Qwen models and Nanonets-OCR2 use Ollama with 4-bit quantized weights, SmolVLM2 uses HuggingFace Transformers with full-precision weights, and Gemini 3 Flash is accessed via API. These differences may affect comparability. All models use the same prompts rather than model-specific tuning. We evaluated only one commercial API model; others (e.g., GPT-4o, Claude) may perform differently.

7. Conclusion

We have presented the first evaluation of VLMs for structured entity extraction from broadcast television chyrons. On the HiChy dataset, locally-deployed Qwen2.5-VL models achieve the best KIE performance (ANLS*≈0.87), outperforming the commercial Gemini 3 Flash API despite its superior raw OCR, while even SmolVLM2 (2.2B) achieves functional KIE (ANLS*=0.745). OCR accuracy and structured extraction capability do not necessarily correlate, suggesting that the KIE task depends more on formatting conventions than on text recognition. Our name origin analysis confirms that the difficulty with Hawaiian content is specifically linguistic, driven by unfamiliarity with Hawaiian name patterns rather than geographic factors, highlighting the need for targeted improvements in how VLMs handle low-resource language orthography. Future work should explore chain-of-thought prompting, model-specific prompt optimization, and fine-tuning on domain-specific data.

8. Acknowledgements

This work was supported by the Andrew W. Mellon Foundation and the CLAMS project at Brandeis University. We thank GBH Archives for providing access to the Hawaiian television broadcast materials used in this study.

9. Bibliographical References

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, et al. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of OCR for low-resource

Task	Error Type	Model	Gold	Predicted
OCR	Case preservation	Gemini Flash	Bobby Bunda / Senate President	BOBBY BUNDA / SENATE PRESIDENT
OCR	Repetition artifact	Qwen 32B	LYNETTE LO TOM / Hawaii Public Television	LYNETTE LO TOM TOM / Hawaii Public Public Television
KIE	Field boundary	Qwen 7B	Name: REP. BRIAN TANIGUCHI (D) Attr: Chairman - House Comm. on Higher Education	Name: REP. BRIAN TANIGUCHI (D) Chairman - House Comm. on Higher Education Attr: *(empty)*
KIE	Background text	Gemini Flash	Name: DAN VUKELICH Attr: Host	Name: DAN VUKELICH Attr: Host / STATE LINE / NEW MEXICO
KIE	Norm. failure	SmolVLM2	Norm: Yamashita-Tungpalan, Eloise	Norm: SEN. ELOISE YAMASHITA-TUNGPALAN
KIE	Hallucination	Nanonets	OCR: GOVERNOR JOHN WAIHEE	OCR: I'm not. I'm not. I'm not. . .

Table 7: Representative error examples. "/" denotes line breaks in OCR output. Gemini's case preservation error is penalized because gold annotations normalize small caps to title case, though the model's reading is arguably correct.

languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Siyu An, Ye Liu, Haoyuan Peng, and Di Yin. 2024. VKIE: The Application of Key Information Extraction on Video Text.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene Text Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas.

Ling Fu et al. 2025. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning.

Google DeepMind. 2025. Introducing Gemini 3: Our most capable models yet.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C.V. Jawahar. 2019. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2023. Watching the News: Towards VideoQA Models that can Read.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. In *Proceedings of the International Conference on Document Analysis and Recognition Workshops (ICDARW)*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s.

Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. Nanonets-OCR-s. Fine-tuned from Qwen2.5-VL-3B-Instruct.

Andrés Marafioti et al. 2025. SmolVLM: Redefining small and efficient multimodal models.

Stephen Mayhew, Terra Blevins, Shuheng Liu, et al. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4322–4337. Association for Computational Linguistics.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In *Workshop on Document Intelligence at NeurIPS*.

David Peer, Philemon Schöpf, Volckmar Nebendahl, Alexander Rietzler, and Sebastian Stabinger. 2024. ANLS* – A Universal Document Processing Metric for Generative Large Language Models.

Qwen Team. 2025. Qwen2.5-VL: To See the World with Wisdom.

Kyeongmin Rim, Owen C. King, Kelley Lynch, Marc Verhagen, and James Pustejovsky. 2025. A Platform for AI-Assisted Archival Metadata Generation. In *Culture and Computing*, pages 183–203. Springer Nature Switzerland.

Brendan Shillingford and Oiwi Parker Jones. 2018. Recovering missing characters in old Hawaiian writing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4929–4934, Brussels, Belgium. Association for Computational Linguistics.

William H. Wilson and Kauanoe Kamanā. 2001. Mai loko mai o ka ’i’ini: Proceeding from a dream: The ’Aha Pūnana Leo connection in Hawaiian language revitalization. In Leanne Hinton and Ken Hale, editors, *The Green Book of Language Revitalization in Practice*, pages 147–176. Academic Press, San Diego, CA.

A. Prompts

All models receive identical prompts for each task. Prompts are delivered via the chat message format with separate system and user roles.

A.1. OCR Prompts

System prompt:

> You are analyzing a single still image from a television program. Your only role is to precisely transcribe the on-screen text. Pay close attention to the exact spelling of words as they appear on the screen. Include every letter, number, and punctuation, as instructed by the user. Do not give any introduction or explanation.

User prompt:

> Transcribe the text in the lower third or chyron area of the screen. Represent line breaks in the on-screen text as \n in your output. Ignore any incidental text in the background. If you find no text, respond with the single word ‘NONE’ and nothing else.

A.2. KIE Prompts

System prompt:

> You are an expert OCR and information extraction system for broadcast television. You extract structured metadata from lower-third chyron graphics. Output only valid JSON. Do not add commentary or explanation.

User prompt (few-shot):

> Extract structured information from the chyron (lower-third graphic) in this image.
>
> Output a JSON object with exactly these three fields:
>
> - `"name-as-written"`: The person’s name exactly as displayed, including titles and designations. Preserve the original capitalization.
> - `"name-normalized"`: The name in “Lastname, Firstname” format. Do not add names or characters not in the original.
> - `"attributes"`: A list of strings, each being a role, title, location, or other characteristic shown in the chyron.
>
> If there is no chyron text, return: `{"name-as-written": "", "name-normalized": "", "attributes": []}`
>
> Examples:
>
> Image shows: “SEN. PATTY MURRAY” and “(D) Washington”
> Output: `{"name-as-written": "SEN. PATTY MURRAY", "name-normalized": "Murray, Patty", "attributes": ["(D) Washington"]}`
>
> Image shows: “Dr. Anthony Fauci” and “Director, NIAID”
> Output: `{"name-as-written": "Dr. Anthony Fauci", "name-normalized": "Fauci, Anthony", "attributes": ["Director, NIAID"]}`
>
> Image shows: “NALUA‘I KAOPUIKI” and “Farmer”
> Output: `{"name-as-written": "NALUA‘I KAOPUIKI", "name-normalized": "Kaopuiki, Nalua‘i", "attributes": ["Farmer"]}`
>
> Now extract from this image:

User prompt (zero-shot): The zero-shot variant uses the same field definitions and empty-response instruction but omits the three examples, ending instead with: “Output only the JSON object, nothing else.”