

Large-scale Multilingual News Image Captioning with LLMs

Yuji Chen¹, Purushoth Velayuthan², Alistair Plum³, Hansi Hettiarachchi¹,
Saroj Basnet⁴, Menan Velayuthan⁵, Marcos Zampieri⁴, Tharindu Ranasinghe¹

¹Lancaster University, UK ²University of Nevada, Reno, USA

³University of Luxembourg, Luxembourg ⁴George Mason University, USA

⁵Utrecht University, Netherlands

t.ranasinghe@lancaster.ac.uk

Abstract

The goal of news image captioning is to generate captions by integrating news article content with corresponding images, highlighting the relationship between textual context and visual elements. The majority of research on news image captioning focuses on English, primarily because datasets in other languages are scarce. To address this limitation, we **create the first large-scale multilingual news image captioning benchmark, LaMuN**, comprising more than 800,000 news images in 59 languages, including several low-resource languages. We further **propose a novel chain-of-thought method** for automatic news image captioning that achieves state-of-the-art results across a diverse set of languages in LaMuN. Together, the benchmark and the proposed method open new avenues for further advances in developing and evaluating multilingual multimodal large language models.

1 Introduction

News image captioning is a variant of image captioning where the goal is to produce entity-rich descriptions that incorporate specific details such as individuals' names and newsworthy events, drawing upon both the visual content and its accompanying article (Rajakumar Kalarani et al., 2023). It has significant practical implications as the accurate captions are a primary means by which screen readers convey visual news content to visually-impaired users (Dognin et al., 2022), and they support content-based indexing and retrieval in digital archives, helping journalists, researchers, and the public locate relevant imagery across vast news collections (Sarto et al., 2022). The rise of vision large language models (VLLMs) has further elevated the importance of news image captioning as an evaluation benchmark, since the task demands joint understanding of textual context and visual content as well as the ability to ground entities and events across the two modalities (Sim et al., 2025).



News Image Caption:
Michelle O'Neill attended the Belfast ceremony alongside Deputy First Minister Emma Little-Pengelly of the Democratic Unionist Party (DUP).

Generic Image Caption:
A crowd of people standing around each other.



News Image Caption:
Maren Mjelde won the Women's Super League in her final season with Chelsea.

Generic Image Caption:
A woman holding a trophy in front of a crowd.

Figure 1: Comparison of news and generic image captions for two different images. The generic image captions were generated by BLIP (Li et al., 2022).

News image captions differ from generic captions in purpose and specificity, as generic captions tend to be purely descriptive, referring to objects in general terms (Figure 1a, 1b). News image captions, are inherently entity-grounded and context-dependent as they situate the image within a broader narrative, naming specific people, places, organisations, and events which are not directly observable from the image itself. Therefore, researchers have introduced different benchmark datasets for news image captioning, such as VISUAL NEWS (Liu et al., 2021), and NY-TIMES800K (Tran et al., 2020). However, these datasets are restricted to English; therefore, current models have been developed and evaluated only in a monolingual English setting.

Recently, Chen et al. (2026) released MUNCHUS, the first multilingual dataset for news image captioning, but it remains limited in both linguistic diversity, covering only nine languages, and source variety, drawing images from a single outlet. Moreover, Chen et al. (2026) evaluate only basic zero-shot and few-shot prompting, which yields consistently poor results, highlighting a research gap for more effective approaches to multilingual news image captioning.

In this paper, we addressed these gaps with the following main contributions:

(i) We **create** the first large-scale multilingual news image captioning benchmark, LaMuN, comprising more than **800,000 news images** from three news sources in **59 languages**.¹

(ii) We **propose** a novel chain-of-thought method for news image captioning that produces state-of-the-art results for all the selected languages in LaMuN.

2 Related Work

Early datasets for news image captioning were developed almost exclusively for English. Ramisa et al. (2017) introduced BREAKINGNEWS, containing around 100,000 news articles with images and captions. Biten et al. (2019) later introduced GOODNEWS, a larger New York Times collection that became a widely used benchmark for the task. This was followed by NYTIMES800K (Tran et al., 2020) and VISUAL NEWS (Liu et al., 2021), the latter containing over one million image–article pairs from multiple news sources. Despite their scale, these benchmarks remain restricted to English. MUNICHUS (Chen et al., 2026) was the first multilingual benchmark, covering nine languages, but its limited linguistic and source diversity leaves multilingual, particularly low-resource, news image captioning largely unexplored.

Models on news image captioning has mainly focused on entity grounding and selecting relevant article context. Biten et al. (2019) dynamically incorporated named entities from articles, while Tran et al. (2020) proposed combining images, faces, objects, and textual context. Later approaches improved context selection and grounding. JOGANIC (Yang et al., 2021) models journalistic caption structure, ICECAP (Hu et al., 2020) concentrates on relevant textual information, and Zhou et al. (2022) use vision–language retrieval to identify useful entities and context. More recently, Qu et al. (2024) combine face–name modelling with CLIP-based retrieval. MERGE (You et al., 2026) further incorporates entity-centric multimodal retrieval. However, these methods are evaluated almost entirely on English benchmarks, leaving their effectiveness across diverse languages largely unknown.

¹Available at <https://huggingface.co/datasets/tharindu/LaMuN>

3 LaMuN: Large-scale Multilingual News Image Captioning Benchmark

#	Language	Code	Train	Test	Unique Articles
1	Albanian	sqi	8,703	1,000	4,592
2	Amharic	amh	12,555	1,000	8,853
3	Arabic	ara	13,003	1,000	6,904
4	Armenian	hye	358	1,000	1,185
5	Bengali	ben	22,686	1,000	8,119
6	Bosnian	bos	10,803	1,000	5,030
7	Bulgarian	bul	9,343	1,000	6,245
8	Burmese	mya	12,492	1,000	7,039
9	Central Kurdish	ckb	4,372	1,000	4,347
10	Chinese (Mandarin)	cmn	58,125	1,000	37,719
11	Croatian	hrv	16,465	1,000	7,265
12	Dari	prs	21,564	1,000	8,707
13	English	eng	135,416	1,000	58,796
14	French	fra	12,381	1,000	4,167
15	Georgian	kat	4,492	1,000	2,861
16	German	deu	29,734	1,000	11,564
17	Gujarati	guj	11,182	1,000	3,192
18	Haitian Creole	hat	883	1,000	1,044
19	Hausa	hau	17,928	1,000	6,483
20	Hindi	hin	14,984	1,000	4,057
21	Igbo	ibo	534	1,000	540
22	Indonesian	ind	13,690	1,000	2,863
23	Japanese	jpn	7,641	1,000	3,805
24	Khmer	khm	9,556	1,000	6,875
25	Kinyarwanda	kin	9,449	1,000	6,483
26	Korean	kor	7,241	1,000	2,624
27	Kurmanji Kurdish	kmr	4,360	1,000	4,339
28	Kyrgyz	kir	4,926	1,000	1,781
29	Lao	lao	7,060	1,000	5,066
30	Lingala	lin	305	1,000	835
31	Macedonian	mkd	2,648	1,000	1,735
32	Marathi	mar	6,918	1,000	1,791
33	Nepali	nep	4,562	1,000	1,983
34	Nigerian Pidgin	pcm	4,293	1,000	2,081
35	North Ndebele	nde	2,454	1,000	2,658
36	Oromo	orm	1,583	1,000	1,294
37	Pashto	pus	10,869	1,000	7,407
38	Persian	fas	64,713	1,000	24,795
39	Polish	pol	7,564	1,000	4,350
40	Punjabi	pan	6,461	1,000	1,520
41	Romanian	ron	5,872	1,000	3,321
42	Russian	rus	23,134	1,000	14,759
43	Scottish Gaelic	gla	0	585	289
44	Serbian	srp	6,626	1,000	1,990
45	Shona	sna	433	1,000	818
46	Sinhala	sin	2,416	1,000	1,046
47	Somali	som	5,312	1,000	4,433
48	Swahili	swa	12,562	1,000	8,163
49	Telugu	tel	3,253	1,000	1,245
50	Thai	tha	21,589	1,000	10,119
51	Tibetan	bod	331	1,000	606
52	Tigrinya	tir	3,028	1,000	2,146
53	Turkish	tur	31,061	1,000	14,953
54	Ukrainian	ukr	9,044	1,000	5,354
55	Urdu	urd	28,707	1,000	11,597
56	Uzbek	uzb	4,340	1,000	1,742
57	Vietnamese	vie	21,881	1,000	15,786
58	Welsh	cym	4,671	1,000	1,889
59	Yoruba	yor	0	295	145
Total			778,556	57,880	383,395

Table 1: Dataset statistics for LaMuN. Languages are categorised as high-resource ■, mid-resource ■, and low-resource ■ following Joshi et al. (2020). Refer to Table 6 for detailed statistics.

We collected news articles and associated images from three major international news sources: BBC News, Voice of America (VOA), and Deutsche

Welle (DW News), spanning from 2001 to December 2025. These sources were selected based on their multilingual coverage, journalistic standards, accessibility, and licensing considerations. To ensure data quality, we first removed images with a height or width of less than 180 pixels. Next, we retained only those examples whose captions contain more than three words. While our experiments utilise only images and articles, the LaMuN also provides additional metadata, such as article titles, which can be used in future studies.

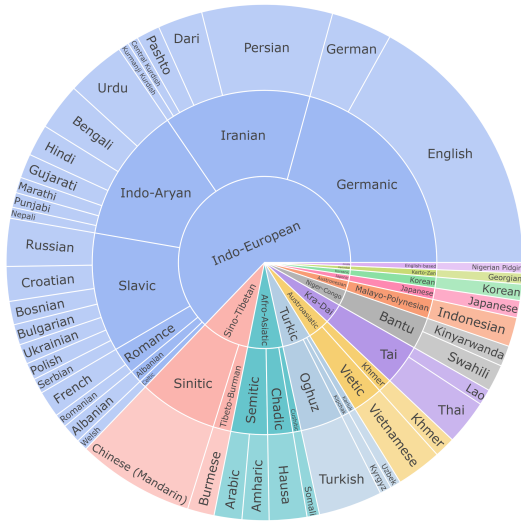


Figure 2: Language family coverage in LaMuN. Segment sizes are proportional to the number of samples per language, and only languages with more than 5,000 samples are shown for clarity.

The resulting LaMuN dataset comprises over 800,000 news images across 59 languages. As shown in Table 1, the final dataset consisted of 778,556 training images and 57,880 test images across 59 languages. According to Joshi et al. (2020), the majority of the languages in LaMuN are mid- and low-resource. Furthermore, as shown in Figure 2, LaMuN covers a diverse range of languages, highlighting its value as a benchmark for multimodal research in low-resource settings.

4 Methods

We experimented with several recently released VLLMs across a selected subset of languages from LaMuN. We evaluate generated captions using two automatic metrics that have been widely adopted in image captioning evaluation: BLEU-4 (Papineni et al., 2002) and CIDEr-R (Oliveira dos Santos et al., 2021). In addition, we report the number of named entities correctly retrieved (F1 score) in the

generated captions as a complementary measure of entity grounding. For this, we used the language specific NER models mentioned in Table 4. We used the exact span matches. We evaluated the following prompting and fine-tuning strategies.

In all strategies, We experiment with four recently released vision-language models: Aya-Vision-8B, Gemma-4-E4B-it, Qwen2.5-VL-7B, and Qwen3-VL-8B.

4.1 Zero-shot Prompting

In the zero-shot setting, we provide the model with the task instruction and the news article context. The prompt specifies the target language and gives guidelines for journalistic captions: keep them concise, identify named entities, maintain factual accuracy, and connect visual content to the news context. The full template is shown in Figure 3.

4.2 Chain of Thought Prompting

We propose a chain-of-thought (CoT) approach that decomposes caption generation into four reasoning steps: (i) visual analysis of the image, (ii) retrieval of the most relevant event context from the article, (iii) extraction of named entities from that context, and (iv) caption synthesis conditioned on all three. We implement two variants, which differ in how the visual analysis step is performed. For both, we use language-specific NER models for each of the nine target languages mentioned in Table 4.

4.2.1 Holistic-CoT

We first use Florence-2 (Xiao et al., 2024) to generate a holistic natural-language description of the image, which serves as a query for retrieving the most event-relevant passage from the article. We embed all article sentences using BGE-M3 (Chen et al., 2024), a multilingual dense retrieval model, and cluster them into event groups with HDBSCAN (McInnes et al., 2017). The cluster whose centroid is most similar to the Florence-2 embedding is selected as the event context, and a language-specific named entity recognition (NER) model extracts a structured set of entities (persons, organisations, locations, dates, and events) from it. The entities and event context are assembled into the structured prompt shown in Figure 4 and then passed, along with the original image, to the LLM.

4.2.2 Spatial-CoT

Holistic-CoT does not convey spatial layout, which can be important for grounding entities to specific

	High-resource									All	Mid-resource									All
	eng	ara	cmn	hin	ind	swa	ben	nep	sin		eng	ara	cmn	hin	ind	swa	ben	nep	sin	
§4.1 Zero-shot Prompting																				
Aya-Vision-8B	2.47	1.77	4.75	2.21	3.08	2.42	1.43	0.36	0.29	2.05	26.12	24.90	10.33	22.45	25.44	25.33	22.11	21.52	12.24	20.94
Gemma-4-E4B-it	2.24	1.71	4.25	1.74	2.35	3.75	1.87	0.51	0.55	2.14	23.54	22.28	9.42	20.22	21.00	24.53	24.02	21.40	16.14	20.41
Qwen2.5-VL-7B	2.94	2.62	4.32	3.19	2.82	4.55	2.53	1.07	0.80	2.72	25.82	21.91	10.05	21.80	23.93	26.71	23.52	21.95	15.12	20.78
Qwen3-VL-8B	2.51	1.79	4.13	3.31	2.71	4.26	2.62	0.78	0.71	2.57	24.03	21.24	10.42	21.13	23.00	25.73	24.38	20.67	15.78	21.31
§4.2.1 Holistic-CoT																				
Aya-Vision-8B	1.88	1.16	4.50	1.93	1.31	3.93	0.51	0.28	0.34	1.75	15.22	8.24	17.92	12.73	12.33	24.55	5.35	5.73	3.82	11.94
Gemma-4-E4B-it	1.95	1.29	4.87	2.05	1.35	4.11	0.54	0.26	0.39	1.83	15.92	8.64	18.31	13.31	12.53	25.74	5.62	5.96	4.23	12.12
Qwen2.5-VL-7B	2.21	1.69	4.95	2.32	1.87	4.45	0.75	0.40	0.45	2.12	17.51	9.82	20.03	14.55	13.70	28.51	6.12	6.55	4.61	13.20
Qwen3-VL-8B	2.03	1.43	4.78	2.14	1.56	4.25	0.65	0.32	0.43	1.92	16.31	9.23	18.72	13.64	12.81	26.32	5.71	6.11	4.31	12.43
§4.2.2 Spatial-CoT																				
Aya-Vision-8B	1.52	0.91	4.72	1.39	1.16	3.38	0.48	0.22	0.40	1.57	15.33	7.12	18.53	11.93	12.63	24.92	6.52	6.73	4.73	11.83
Gemma-4-E4B-it	1.63	0.94	5.12	1.47	1.23	3.54	0.56	0.23	0.38	1.65	15.64	7.51	19.52	12.05	13.61	25.82	6.85	6.71	5.38	12.51
Qwen2.5-VL-7B	1.89	1.23	5.36	1.76	1.59	3.98	0.71	0.31	0.44	1.93	18.32	9.11	22.73	14.61	15.43	30.03	7.94	8.14	6.21	14.82
Qwen3-VL-8B	1.75	1.05	5.08	1.62	1.36	3.62	0.60	0.26	0.45	1.72	16.54	8.04	20.63	13.31	14.41	27.13	7.22	7.43	5.23	13.53
§4.3 Basic Finetuning																				
Aya-Vision-8B	5.14	4.01	6.83	3.91	3.88	5.15	2.29	1.01	1.08	3.69	49.13	40.72	63.32	38.55	37.16	51.83	18.67	9.33	9.75	35.21
Gemma-4-E4B-it	4.87	3.79	6.52	3.66	3.50	4.98	2.07	0.95	0.90	3.47	45.82	38.64	59.24	35.55	33.13	49.14	16.95	8.44	8.32	32.93
Qwen2.5-VL-7B	5.70	4.32	7.22	4.39	4.31	5.62	2.85	1.48	1.42	4.14	54.74	44.44	69.41	43.76	41.56	56.62	22.44	12.36	11.61	39.55
Qwen3-VL-8B	5.41	4.18	6.99	4.15	4.02	5.48	2.55	1.32	1.29	3.97	52.14	43.64	66.62	41.93	39.23	55.45	20.64	11.22	10.69	38.12
§4.3 Holistic-CoT Finetuning																				
Aya-Vision-8B	6.28	4.92	8.05	4.77	4.65	6.22	3.25	1.64	1.61	4.57	65.92	54.12	82.33	51.54	50.54	67.73	28.82	15.64	15.34	48.82
Gemma-4-E4B-it	5.93	4.68	7.59	4.58	4.38	6.04	2.84	1.41	1.38	4.41	62.02	51.51	77.86	49.33	46.82	65.12	25.44	13.75	12.83	45.93
Qwen2.5-VL-7B	7.08	5.65	8.84	5.58	5.53	6.98	3.79	2.02	2.01	5.32	74.51	61.82	91.92	60.11	58.24	76.14	35.12	19.42	19.31	55.92
Qwen3-VL-8B	6.62	5.36	8.32	5.38	5.12	6.63	3.40	1.86	1.78	5.07	69.82	58.82	87.13	56.55	53.85	72.53	31.52	17.55	16.24	53.26
§4.3 Spatial-CoT Finetuning																				
Aya-Vision-8B	7.05	5.68	8.81	5.44	5.31	7.04	3.64	1.92	1.96	5.26	74.24	62.52	92.22	58.78	57.64	76.44	32.85	18.23	17.51	54.33
Gemma-4-E4B-it	6.68	5.32	8.41	5.18	5.02	6.61	3.22	1.68	1.63	4.95	69.15	58.24	86.54	55.64	52.34	72.83	29.64	16.25	14.83	51.23
Qwen2.5-VL-7B	7.95	6.51	9.77	6.42	6.25	7.93	4.31	2.52	2.43	6.24	81.92	68.78	100.52	66.82	64.14	83.92	39.31	22.91	21.62	63.12
Qwen3-VL-8B	7.52	6.08	9.31	6.05	5.84	7.48	3.98	2.18	2.12	5.82	78.13	65.52	96.24	63.42	61.24	80.23	36.24	20.42	19.81	59.84

Table 2: BLEU-4 (left-side) (Papineni et al., 2002) and CIDEr-R (right-side) (Oliveira dos Santos et al., 2021) evaluation scores across test sets in selected languages in LaMuN. The best result for each language (any method) is in **bold**.

visual regions. Spatial-CoT instead segments the image into foreground and background regions using SAM3 (Carion et al., 2025), the third generation of the Segment Anything Model. Each region is assigned a semantic label (e.g., *person*, *building*), and NER entities are distributed by type: person entities to *person* regions, location and organisation entities to *building* regions. The model receives the original image, the cropped region images, an event context retrieved via the same BGE-M3 process as Holistic-CoT, and an entity-to-region assignment list. The prompt instructs it to describe the scene progressively from foreground to background, following the region order, and to use the assigned entity names exactly.

4.3 Basic Instruction Fine-tuning

To assess whether task-specific adaptation closes the gap between zero-shot performance and the reference captions, we also performed instruction fine-tuning. For each target language, we constructed training examples using the same prompt template as the zero-shot setting. We fine-tune LLMs using LoRA (Hu et al., 2022) for each language and evaluate them on the test set using the same prompt.

4.4 CoT Fine-tuning

To test whether the gains from CoT prompting persist under adaptation, we additionally fine-tune on the structured CoT prompts rather than the zero-shot template. We fine-tune the LLMs using LoRA (Hu et al., 2022) separately for each language and evaluate them on the test set using the same prompt. The training hyperparameters are provided in Table 5 in the Appendix.

5 Results and Discussion

Tables 2 and 3 report BLEU-4, CIDEr-R, NER F1 scores for generated captions across nine selected languages in LaMuN. We describe our main findings below.

(i) **Chain-of-Thought Prompting and Fine-tuning** - Neither Holistic-CoT nor Spatial-CoT outperforms zero-shot prompting when used in a prompting setting. Across all languages and metrics, zero-shot consistently yields higher scores than both CoT variants. We attribute this to the additional complexity introduced by multi-step reasoning, which smaller instruction-tuned models struggle to follow reliably without task-specific adaptation. However, fine-tuning on structured CoT prompts yields substantial gains, as both CoT

	High-resource			Mid-resource			Low-resource			All
	eng	ara	cmn	hin	ind	swa	ben	nep	sin	
§4.1 Zero-shot Prompting										
Aya-Vision-8B	6.23	5.12	7.45	8.12	4.23	11.34	2.45	3.67	1.89	5.61
Gemma-4-E4B-1t	5.89	4.78	7.01	7.56	3.98	10.67	2.12	3.23	1.67	5.21
Qwen2.5-VL-7B	7.12	5.89	8.34	9.23	4.89	13.45	2.89	4.23	2.12	6.46
Qwen3-VL-8B	7.45	6.12	8.67	9.56	5.12	13.89	3.12	4.56	2.34	6.76
§4.2.1 Holistic-CoT										
Aya-Vision-8B	12.89	10.89	15.05	16.58	8.42	22.30	4.89	8.03	3.86	11.43
Gemma-4-E4B-1t	14.23	12.03	16.62	18.31	9.30	24.63	5.40	8.87	4.27	12.63
Qwen2.5-VL-7B	15.67	13.23	18.28	20.14	10.23	27.09	5.94	9.76	4.70	13.89
Qwen3-VL-8B	16.23	13.78	18.95	20.87	10.67	28.12	6.23	10.12	4.92	14.43
§4.2.2 Spatial-CoT										
Aya-Vision-8B	12.81	10.46	14.71	15.43	7.99	22.00	5.81	10.01	4.22	11.49
Gemma-4-E4B-1t	14.16	11.56	16.25	17.05	8.83	24.29	6.42	11.05	4.66	12.70
Qwen2.5-VL-7B	15.58	12.72	17.88	18.76	9.71	26.72	7.06	12.16	5.13	13.97
Qwen3-VL-8B	16.12	13.23	18.52	19.43	10.12	27.73	7.39	12.61	5.37	14.50
§4.3 Basic Finetuning										
Aya-Vision-8B	18.34	15.67	21.23	23.45	12.34	31.23	7.89	11.23	5.67	16.34
Gemma-4-E4B-1t	17.23	14.89	20.12	22.34	11.56	29.89	7.23	10.45	5.23	15.44
Qwen2.5-VL-7B	20.45	17.34	23.56	26.78	13.89	34.56	8.89	12.89	6.45	18.31
Qwen3-VL-8B	21.23	18.12	24.34	27.89	14.34	35.89	9.23	13.34	6.78	19.02
§4.3 Holistic-CoT Finetuning										
Aya-Vision-8B	35.67	31.23	41.23	45.67	25.67	58.34	15.67	21.34	11.23	31.78
Gemma-4-E4B-1t	33.45	29.67	39.45	43.89	24.34	56.23	14.56	19.89	10.45	30.21
Qwen2.5-VL-7B	38.23	33.89	44.23	49.12	27.89	62.34	17.23	23.45	12.34	34.30
Qwen3-VL-8B	39.45	35.12	45.67	50.89	28.89	64.23	17.89	24.34	12.89	35.49
§4.3 Spatial-CoT Finetuning										
Aya-Vision-8B	36.89	32.45	42.67	47.23	26.89	60.12	16.89	23.12	12.34	33.18
Gemma-4-E4B-1t	34.67	30.89	40.89	45.34	25.56	58.01	15.78	21.56	11.56	31.58
Qwen2.5-VL-7B	39.89	35.34	46.12	51.34	29.12	64.89	18.56	25.34	13.45	36.02
Qwen3-VL-8B	41.23	36.78	47.56	53.12	30.23	67.12	19.34	26.45	14.12	37.33

Table 3: Entity grounding (NER F1) scores between generated caption and the human caption across test sets in selected languages in LaMuN. The best result for each language (any method) is in **bold**.

fine-tuning strategies outperform basic fine-tuning across all languages. This confirms that although models cannot reliably follow CoT reasoning out of the box, they benefit significantly from exposure to the same reasoning structure during training. We also observe that Spatial-CoT provides slightly better results than Holistic-CoT.

(ii) **Performance Across Resource Tiers** - High-resource languages (English, Arabic, Chinese) yield the strongest results across all metrics. For CIDEr-R, the best-performing configuration (Qwen2.5-VL-7B with Spatial-CoT fine-tuning) achieves 81.88, 68.77, and 100.46 for English, Arabic, and Chinese, respectively. Mid-resource languages (Hindi, Indonesian, Swahili) show competitive performance, with Swahili reaching 84.06. Low-resource languages (Bengali, Nepali, Sinhala) remain challenging, with Sinhala peaking at only 21.84 CIDEr-R and 14.12 NER F1. We attribute that news image captioning is a challenging task for low-resource languages.

(iii) **Metric Alignment and Divergence** - While BLEU-4 and CIDEr-R follow similar trends across configurations, NER F1 reveals a different finding. The gap between basic fine-tuning and Spatial-CoT fine-tuning is more pronounced in NER F1 for low-resource languages, suggesting that CoT fine-tuning particularly benefits entity grounding.

(iv) **Model Ranking** - Across most settings, Qwen2.5-VL-7B slightly outperforms other models. The gap is most noticeable in low-resource languages. However, Qwen3-VL-8B performs better than Qwen2.5-VL-7B in entity grounding suggesting that there is no clear winner model. Furthermore, model rankings on high-resource languages do not reliably predict rankings on low-resource languages, reinforcing the need for diverse multilingual evaluation.

6 Conclusion and Future Work

In this paper, we introduced LaMuN, the first large-scale multilingual benchmark for news image captioning comprising more than 800,000 news images across 59 languages and three international news sources. We also proposed two chain-of-thought approaches, Holistic-CoT and Spatial-CoT, which incorporate visual information, relevant article context, and named entities into the caption generation process. Our experiments show that chain-of-thought reasoning does not improve performance when used only as a prompting strategy. However, both approaches yield substantial gains when models are fine-tuned using the same structured prompts. Spatial-CoT fine-tuning performs best overall and provides particularly strong improvements in entity grounding.

Our results also show that news image captioning remains challenging for low-resource languages, where performance is considerably lower than for high-resource languages. Furthermore, model rankings on high-resource languages do not always transfer to low-resource settings, highlighting the importance of multilingual evaluation. We release LaMuN to support further research on multilingual news image captioning and multimodal language models.

In future work we would like to explore the performance of the evaluated models on articles containing non-literal language, such as humour and sarcasm, a known challenge for VLMs (Farabi et al., 2024; Basnet et al., 2025). Furthermore, we would like to further explore their performance on the existing low-resource languages in LaMuN. Due to space limitations, we were unable to carry out a more detailed evaluation of the models’ performance on this set of languages. Finally, we plan to extend evaluation to more languages in LaMuN and investigate stronger multilingual LLMs and alternative grounding strategies, particularly for low-resource languages.

Acknowledgements

The experiments in this paper were conducted on the High End Computing (HEC) Cluster at Lancaster University, which is funded through a combination of central funding and contributions from individual research grants.

This work is funded by COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology).

Limitations

We acknowledge the following limitations.

Evaluation coverage. Although LaMuN spans 59 languages, our experiments evaluate captioning performance on a subset of nine languages, selected to span the high-, mid-, and low-resource tiers and to cover languages for which reliable language-specific named entity recognition (NER) models are available. The page limit constraints does not let us evaluate all 59 languages. The behaviour of the proposed methods on the remaining languages, and on writing systems or language families not represented in our nine-language subset, therefore remains to be explored. We release the full benchmark precisely so that future work can extend evaluation to the languages we do not study here.

Automatic evaluation only. We assess generated captions using BLEU-4, CIDEr-R, and NER F1. These automatic metrics correlate only imperfectly with human judgements of caption quality, and n-gram overlap metrics in particular can penalise fluent, factually correct captions that differ lexically from the reference. We do not conduct a human evaluation as it is difficult to find human annotators for a large scale multilingual setting.

Model and scale constraints. We experiment with a fixed set of small open-weight vision-language models and adapt them with LoRA rather than full fine-tuning, owing to computational constraints. We do not evaluate larger or proprietary models, and our finding that chain-of-thought reasoning helps only under fine-tuning, not as a prompting strategy, may not transfer to substantially larger models with stronger instruction-following abilities.

Ethics Statement

LaMuN was collected from publicly available news articles and images published by BBC News, Voice of America (VOA), and Deutsche Welle (DW), and none of the records were edited during collection. The content is publicly accessible, and our collection process adhered to each source’s terms of service and respected the access restrictions specified in their robots.txt directives. We provide the URL to the original news article, ensuring proper attribution and enabling verification of the source material. We distribute the dataset in a form that preserves attribution to the original sources and prohibits alteration of the released instances, and we release it strictly for non-commercial research purposes; users are responsible for complying with the terms of the respective news providers when using the underlying content.

References

- Saroj Basnet, Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, and Marcos Zampieri. 2025. Evaluating open-source vision-language models for multimodal sarcasm detection. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1442–1447. IEEE.
- Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. [Good news, everyone! context driven entity-aware captioning for news images](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, and 19 others. 2025. [SAM 3: Segment anything with concepts](#). *Preprint*, arXiv:2511.16719.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Yuji Chen, Alistair Plum, Hansi Hettiarachchi, Diptesh Kanojia, Saroj Basnet, Marcos Zampieri, and Tharindu Ranasinghe. 2026. MUNIChus: Multilingual News Image Captioning Benchmark. In *Proceedings of the Fifteenth Language Resources and*

- Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).
- Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. [Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge](#). *J. Artif. Int. Res.*, 73.
- Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8020–8028.
- Anwen Hu, Shizhe Chen, and Qin Jin. 2020. [Icecap: Information concentrated entity-aware image captioning](#). In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4217–4225, New York, NY, USA. Association for Computing Machinery.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. [Visual news: Benchmark and challenges in news image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.
- Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. [CIDEr-R: Robust consensus-based image description evaluation](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2024. [Visually-aware context modeling for news image captioning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2927–2943, Mexico City, Mexico. Association for Computational Linguistics.
- Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, Niyati Chhaya, and Sumit Shekhar. 2023. [“let’s not quote out of context”: Unified vision-language pre-training for context assisted image captioning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 695–706, Toronto, Canada. Association for Computational Linguistics.
- Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2017. [The BreakingNews dataset](#). In *Proceedings of the Sixth Workshop on Vision and Language*, pages 38–39, Valencia, Spain. Association for Computational Linguistics.
- Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. [Retrieval-augmented transformer for image captioning](#). In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing, CBMI '22*, page 1–7, New York, NY, USA. Association for Computing Machinery.
- Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoan Fang. 2025. [Can VLMs actually see and read? a survey on modality collapse in vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.
- Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. [Transform and tell: Entity-aware news image captioning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13032–13042.
- Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.
- Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021. [Journalistic guidelines aware news image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoxing You, Qiang Huang, Lingyu Li, Chi Zhang, Xiaopeng Liu, Min Zhang, and Jun Yu. 2026. Knowledge completes the vision: A multimodal entity-aware retrieval-augmented GEneration framework for news image captioning. volume 40, pages 12108–12116.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022. [Focus! relevant and sufficient context selection for news image captioning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

Language	Model
English	dslim/roberta-large-NER
Arabic	aubmindlab/bert-base-arabertv02-ner
Hindi	MichaelHuang/muril_base_cased_hindi_ner
Indonesian	cahya/xlm-roberta-base-indonesian-NER
Swahili	Davlan/xlm-roberta-large-masakhaner
Sinhala	asanka25/xlm-roberta-base-finetuned-conll03-english-finetuned-sinhala
Chinese	shibing624/macbert4ner-base-chinese
Bengali	sagorsarker/mbert-bengali-ner
Nepali	Saugatkafley/mbert-Nepali-NER

Table 4: Named entity recognition models used for each language.

Hyperparameter	Setting
Base models	Aya-Vision-8B, Gemma-4-E4B-it, Qwen2.5-VL-7B, Qwen3-VL-8B
Adaptation method	LoRA (Hu et al., 2022)
LoRA rank (r)	16
LoRA α	32
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj
Optimiser	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight decay	0.01
Learning rate	2×10^{-4}
LR schedule	Cosine decay with 3% linear warmup
Epochs	2
Per-device batch size	4
Gradient accumulation steps	4
Effective batch size	16
Maximum sequence length	8192
Precision	bf16
Hardware	1 $\times$ NVIDIA H200 (Hopper, 141 GB)

Table 5: Hyperparameters used for instruction fine-tuning. The same configuration is used across models and languages.

A Prompt

Prompt template used for zero-shot image captioning is shown in Figure 3.

B Training Hyperparameters

Table 5 summarises the hyperparameters used for instruction fine-tuning. We use the same training configuration across all models and languages.

C Methodology

C.1 Chain of Thought Prompting

A central challenge in news image captioning is that the important information needed to produce an accurate caption the people, organisations, locations, and events depicted is rarely recoverable from the image pixels alone. It must instead be grounded in the accompanying article. Directly

You are writing a caption for a newspaper image.

Given the image and this news article excerpt:

[News content]

Task: Write a concise, informative caption for this image in {language}.

Guidelines:

- Write in {language} language only
 - Keep it brief
 - Identify and include: people’s names, locations, and organisations
 - Connect what you see in the image to the news context
 - Use journalistic style (factual, clear, objective)
 - Focus on the main subject of the image
- Caption in {language}:

Figure 3: Prompt template used for zero-shot image captioning. The {language} placeholder is replaced with the target language name (e.g., “English”, “Arabic”) at inference time.

prompting a large language model (LLM) with the image and the full article conflates this grounding problem with the surface task of caption generation, and in practice leads to hallucinated entities and generic descriptions that omit the named entities readers expect. To address this, we propose a chain-of-thought (CoT) approach that explicitly decomposes caption generation into four interpretable reasoning steps: (i) visual analysis of the image to establish what is depicted, (ii) retrieval of the most relevant event context from the article, (iii) extraction of named entities from that context, and (iv) caption synthesis conditioned jointly on the image, the retrieved context, and the extracted entities. Decomposing the task in this way isolates each sub-problem, allowing the final synthesis step to focus on fluent generation while the grounding burden is handled by dedicated, more reliable components.

We implement two variants of this approach, which differ only in how the visual analysis step (i) is performed: Holistic-CoT produces a single global description of the image, whereas Spatial-CoT additionally decomposes the image into regions and grounds entities to them. The remaining steps retrieval, NER, and synthesis are shared between the two variants. For both, we use language-specific NER models for each of the nine target languages rather than a single multilingual model, as language-specific models gave consistently more reliable entity extraction in our preliminary experi-

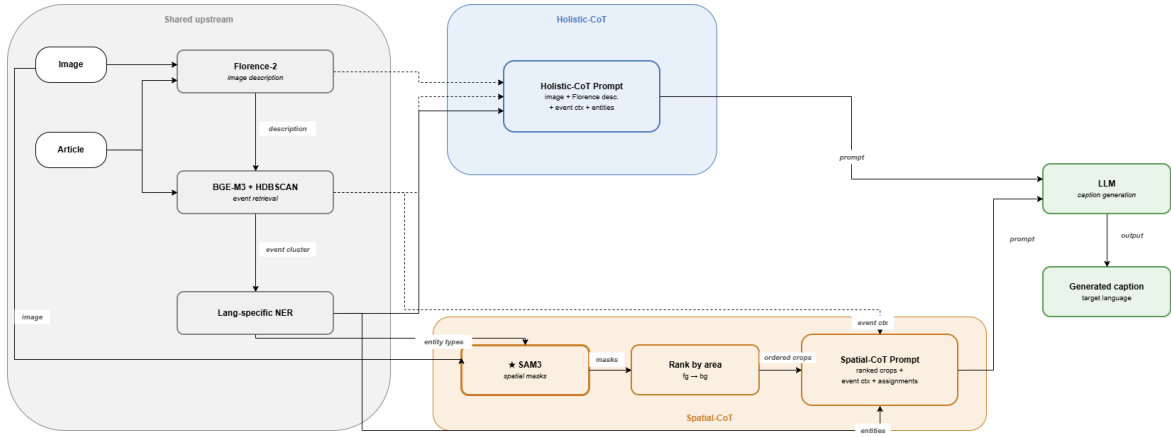


Figure 4: Overview of CoT variants.

ments; the full list of models is provided in Table 4. We used the most popular NER model per language as it appears in HuggingFace.

C.1.1 Holistic-CoT

In Holistic-CoT, we first use Florence-2 (Xiao et al., 2024) to generate a holistic natural-language description of the image. This description acts as a language-space proxy for the visual content, enabling us to bridge the image and the textual article within a shared representation. We use this description as a query for retrieving the most event-relevant passage from the article. Specifically, we embed all article sentences using BGE-M3 (Chen et al., 2024), a multilingual dense retrieval model chosen for its strong cross-lingual performance across our nine target languages, and cluster them into candidate event groups using HDBSCAN (McInnes et al., 2017). We use HDBSCAN rather than a fixed- k clustering method because the number of distinct events discussed in an article is not known *a priori* and varies considerably across articles; HDBSCAN infers the number of clusters from the data and is robust to outlier sentences that do not belong to any event. The cluster whose centroid is most similar (by cosine similarity) to the Florence-2 description embedding is selected as the event context. A language-specific NER model then extracts a structured set of entities persons, organisations, locations, dates, and events from this context. Finally, the extracted entities and the event context are assembled into the structured prompt shown in Figure 4, which is passed together with the original image to the LLM for caption synthesis.

C.1.2 Spatial-CoT

While Holistic-CoT supplies the LLM with the relevant entities and event context, its global image description does not convey the *spatial layout* of the scene. This can be important for correctly grounding entities to specific visual regions for example, attributing the right name to the right person when several individuals appear in a single image. Spatial-CoT addresses this by replacing the holistic visual analysis with a region-level decomposition. We segment the image into foreground and background regions using SAM3 (Carion et al., 2025), the third generation of the Segment Anything Model, and assign each region a semantic label (e.g., *person*, *building*). The NER entities are then distributed across regions by type: person entities are associated with *person* regions, while location and organisation entities are associated with *building* regions. This type-based heuristic provides a soft alignment between textual entities and visual regions without requiring explicit visual grounding supervision. The LLM receives the original image, the cropped region images, an event context retrieved via the same BGE-M3 process used in Holistic-CoT, and the resulting entity-to-region assignment list. The prompt instructs the model to describe the scene progressively from foreground to background, following the region ordering, and to use the assigned entity names verbatim. This ordered, region-by-region formulation encourages the model to produce captions in which named entities are explicitly tied to the visual elements they describe, rather than mentioned in isolation.

#	Language	Code	Family	[Train]	[Test]	Unique Articles	Avg Images / Article	Avg Content Tokens	Avg Caption Tokens	News Sources
1	Albanian	sqi	Indo-European (Albanian)	8,703	1,000	4592	2.11	617.21	8.67	DW
2	Amharic	amh	Afro-Asiatic (Semitic)	12,555	1,000	8853	1.53	448.03	10.22	BBC DW VOA
3	Arabic	ara	Afro-Asiatic (Semitic)	13,003	1,000	6904	2.03	814.39	13.28	BBC DW
4	Armenian	hye	Indo-European (Armenian)	358	1,000	1185	1.15	966.67	8.27	VOA
5	Bengali	ben	Indo-European (Indo-Aryan)	22,686	1,000	8119	2.92	1052.13	10.14	BBC DW VOA
6	Bosnian	bos	Indo-European (Slavic)	10,803	1,000	5030	2.35	802.34	9.7	DW
7	Bulgarian	bul	Indo-European (Slavic)	9,343	1,000	6245	1.66	627.57	9.06	DW
8	Burmese	mya	Sino-Tibetan (Tibeto-Burman)	12,492	1,000	7039	1.92	1863.94	15.76	BBC VOA
9	Central Kurdish	ckb	Indo-European (Iranian)	4,372	1,000	4347	1.24	785.32	5.4	VOA
10	Chinese (Mandarin)	cmn	Sino-Tibetan (Sinitic)	58,125	1,000	37719	1.57	2150.15	20.79	BBC VOA
11	Croatian	hrv	Indo-European (Slavic)	16,465	1,000	7265	2.4	726.34	10.21	DW
12	Dari	prs	Indo-European (Iranian)	21,564	1,000	8707	2.59	479.63	16.53	BBC DW VOA
13	English	eng	Indo-European (Germanic)	135,416	1,000	58796	2.32	530.79	15.08	BBC DW VOA
14	French	fra	Indo-European (Romance)	12,381	1,000	4167	3.21	1326.02	17.03	BBC DW
15	Georgian	kat	Kartvelian (Karto-Zan)	4,492	1,000	2861	1.92	787.08	7.22	VOA
16	German	deu	Indo-European (Germanic)	29,734	1,000	11564	2.66	765.66	11.85	DW
17	Gujarati	guj	Indo-European (Indo-Aryan)	11,182	1,000	3192	3.82	1227.42	10.15	BBC
18	Haitian Creole	hat	Creole (French-based)	883	1,000	1044	1.8	178.82	14.64	VOA
19	Hausa	hau	Afro-Asiatic (Chadic)	17,928	1,000	6483	2.92	361.48	12.54	BBC DW VOA
20	Hindi	hin	Indo-European (Indo-Aryan)	14,984	1,000	4057	3.94	1786.29	12.54	BBC DW
21	Igbo	ibo	Niger-Congo (Volta-Niger)	534	1,000	540	2.84	709.64	11.09	BBC
22	Indonesian	ind	Austronesian (Malayo-Polynesian)	13,690	1,000	2863	5.13	1671.56	15.77	BBC DW
23	Japanese	jpn	Japonic (Japanese)	7,641	1,000	3805	2.27	1286.89	20.63	BBC
24	Khmer	khm	Austroasiatic (Khmer)	9,556	1,000	6875	1.54	9494.9	31.21	VOA
25	Kinyarwanda	kin	Niger-Congo (Bantu)	9,449	1,000	6483	1.61	545.77	12.36	BBC VOA
26	Korean	kor	Koreanic (Korean)	7,241	1,000	2624	3.14	1777.02	24.68	BBC
27	Kurmanji Kurdish	kmr	Indo-European (Iranian)	4,360	1,000	4339	1.24	785.22	5.39	VOA
28	Kyrgyz	kir	Turkic (Kipchak)	4,926	1,000	1781	3.33	911.4	9.14	BBC
29	Lao	lao	Kra-Dai (Tai)	7,060	1,000	5066	1.59	1983.87	35.12	VOA
30	Lingala	lin	Niger-Congo (Bantu)	305	1,000	835	1.56	214.06	24.87	VOA
31	Macedonian	mkd	Indo-European (Slavic)	2,648	1,000	1735	2.1	676.88	12.3	DW
32	Marathi	mar	Indo-European (Indo-Aryan)	6,918	1,000	1791	4.42	1269.09	10.25	BBC
33	Nepali	nep	Indo-European (Indo-Aryan)	4,562	1,000	1983	2.8	878.34	10.66	BBC
34	Nigerian Pidgin	pcm	Creole (English-based)	4,293	1,000	2081	2.54	949.04	13.72	BBC
35	North Ndebele	nde	Niger-Congo (Bantu)	2,454	1,000	2658	1.3	481.72	7	VOA
36	Oromo	orm	Afro-Asiatic (Cushitic)	1,583	1,000	1294	2	875.75	10.18	BBC
37	Pashto	pus	Indo-European (Iranian)	10,869	1,000	7407	1.6	962.56	16.36	BBC DW VOA
38	Persian	fas	Indo-European (Iranian)	64,713	1,000	24795	2.65	2149.68	16.15	BBC DW VOA
39	Polish	pol	Indo-European (Slavic)	7,564	1,000	4350	1.97	717.68	10.19	BBC DW
40	Punjabi	pan	Indo-European (Indo-Aryan)	6,461	1,000	1520	4.91	1411.16	14.16	BBC
41	Romanian	ron	Indo-European (Romance)	5,872	1,000	3321	2.07	835.2	10.48	DW
42	Russian	rus	Indo-European (Slavic)	23,134	1,000	14759	1.64	1280.46	10.37	BBC DW VOA
43	Scottish Gaelic	gla	Indo-European (Celtic)	0	585	289	2.02	560.22	14.88	BBC
44	Serbian	srp	Indo-European (Slavic)	6,626	1,000	1990	3.83	1344.64	12.02	BBC
45	Shona	sna	Niger-Congo (Bantu)	433	1,000	818	1.75	529.34	6.09	VOA
46	Sinhala	sin	Indo-European (Indo-Aryan)	2,416	1,000	1046	3.27	1194.25	16.39	BBC
47	Somali	som	Afro-Asiatic (Cushitic)	5,312	1,000	4433	1.42	518.21	7.95	BBC VOA
48	Swahili	swa	Niger-Congo (Bantu)	12,562	1,000	8163	1.66	505.26	13.24	BBC DW VOA
49	Telugu	tel	Dravidian (South-Central)	3,253	1,000	1245	3.42	730.93	7.54	BBC
50	Thai	tha	Kra-Dai (Tai)	21,589	1,000	10119	2.23	1436.51	20.87	BBC VOA
51	Tibetan	bod	Sino-Tibetan (Tibeto-Burman)	331	1,000	606	2.2	2248.97	14.44	VOA
52	Tigrinya	tir	Afro-Asiatic (Semitic)	3,028	1,000	2146	1.88	732.16	9.79	BBC VOA
53	Turkish	tur	Turkic (Oghuz)	31,061	1,000	14953	2.14	740.71	10.9	BBC DW VOA
54	Ukrainian	ukr	Indo-European (Slavic)	9,044	1,000	5354	1.88	799.28	9.9	BBC DW
55	Urdu	urd	Indo-European (Indo-Aryan)	28,707	1,000	11597	2.56	4575.03	17.24	BBC DW VOA
56	Uzbek	uzb	Turkic (Karluk)	4,340	1,000	1742	3.07	1291.93	12.66	BBC
57	Vietnamese	vie	Austroasiatic (Vietic)	21,881	1,000	15786	1.45	1829.61	19.38	BBC VOA
58	Welsh	cym	Indo-European (Celtic)	4,671	1,000	1889	3	672.81	14.81	BBC
59	Yoruba	yor	Niger-Congo (Volta-Niger)	0	295	145	2.03	607.95	13.61	BBC

Table 6: Dataset statistics for LaMuN. Languages are categorised as high-resource ■, mid-resource ■, and low-resource ■ following Joshi et al. (2020).