JHU's Submission to the AmericasNLP 2025 Shared Task on the Creation of Educational Materials for Indigenous Languages

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, David Yarowsky
Center for Language and Speech Processing,
Johns Hopkins University
{tlupick1,ls1,kchapar1,yarowsky}@jhu.edu

Abstract

This paper presents JHU's submission to the AmericasNLP shared task on the creation of educational materials for Indigenous languages. The task involves transforming a base sentence given one or more tags that correspond to grammatical features, such as negation or tense. The task also spans four languages: Bribri, Maya, Guaraní, and Nahuatl. We experiment with augmenting prompts to large language models with different information, chain of thought prompting, ensembling large language models by majority voting, and training a pointer-generator network. Our System 1, an ensemble of large language models, achieves the best performance on Maya and Guaraní, building upon the previous successes in leveraging large language models for this task and highlighting the effectiveness of ensembling large language models.

1 Introduction

The AmericasNLP 2025 shared task on the creation of educational materials (de Gibert et al., 2025) proposes automated generation of educational materials for low-resource Indigenous languages in the Americas. Many of these languages are endangered, with few remaining speakers, and lack the large datasets necessary to leverage advances in Natural Language Processing (NLP) as languages like English and Spanish do. The shared task challenged teams to develop NLP systems to create educational exercises for Bribri, Maya, Guaraní, and Nahuatl. These exercises involve applying grammatical transformations—such as tense changes or negation—to base sentences.

Each team received a limited training dataset for each language. This dataset contained base sentences, the corresponding grammatical modifications, and the correctly transformed output sentences. Using this data, teams were expected to develop NLP systems which, given a base sentence and a grammar modification, could produce the correctly modified output sentence.

By leveraging NLP to generate grammatical exercises, this task intends to reduce the burden on the small number of fluent speakers in these languages who would otherwise need to manually develop learning resources. This automation can enable communities to create a broader range of instructional materials with less effort, making language learning more accessible.

Our approach is based on an ensemble of several distinct methods, including novel extensions on the large language model (LLM) methods successfully deployed by top performing systems of the 2024 shared task (Vasselli et al., 2024; Bui and von der Wense, 2024; Haley, 2024), combined with additional components including linguistic information specific to each language, part-of-speech tagging, chain-of-thought reasoning, and model ensembling using majority voting. We additionally train a pointer-generator LSTM leveraging additional Bribri data. Our ensemble system, using majority voting from LLM outputs generated with varying prompt configurations, achieves the highest performance on Maya and Guarani compared to other teams. We release our code on GitHub[1].

2 Data

2.1 Task Data

The task provided training, development, and test data in Bribri, Maya, Guaraní, and Nahuatl. Each data split contained base sentences, the change to apply to each base sentence. The training and development data additionally contain the correctly transformed base sentence. The training data includes 309, 584, 178, and 392 examples for Bribri, Maya, Guaraní, and Nahuatl respectively. The development data includes 212 Bribri examples,

[1] https://github.com/KentonMurray/AmericasNLP2025

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 105–111

149 Maya examples, 79 Guaraní examples, and 176 Nahuatl examples. The test data includes 480 Bribri examples, 310 Maya examples, 364 Guaraní examples, and 120 Nahuatl examples.

2.2 Additional Bribri Data

For one of our submitted systems, the pointer-generator network, we create additional training data by extracting verb conjugation tables from *Gramática de la lengua bribri*, a Bribri reference grammar (Murillo, 2018). From this process, we extracted 482 unique verbs and constructed 1400 additional single-verb training examples.

3 Methods

3.1 LLM Prompting

We conduct few-shot prompting experiments utilizing variations of the prompt in Table 1, modified from the prompt used in the 2024 submission to this task by the JAJ team (Vasselli et al., 2024). This prompt provides a well structured format that allows us to experiment with the inclusion of additional information, namely part of speech tags and grammar information from a reference book. An explicit system instruction to output only the target sentence is included as initial testing showed that with such an instruction, outputs were inconsistently formatted and occasionally multiple hypotheses for target sentences were generated. We also observed that the LLMs would sometimes first generate reasoning text, particularly when including fewer few-shot examples in the prompt.

Examples are to include in the prompt are chosen in the following manner: Given a maximum number of examples to include and a test example with n change tags, we first select all examples from the training data such that all n tags in the test example match those in the training examples, then sort in descending order by combined BLEU (Papineni et al., 2002) and chrF (Popović, 2015) score and select up to the given maximum number of examples. If more examples are needed, we select additional training examples which overlap with n - 1 of the change tags in the test example, then again sort and select the top examples by combined BLEU and chrF. If more examples are still needed, we continue this process down to an overlap of 1 change tag.

For this few-shot prompting approach, we experiment with including a maximum of 3, 5, 10, and 20 examples.

```
SYSTEM:
You are a helpful assistant with expertise
in linguistics. Output only the target
sentence in your response with no
additional punctuation.

USER:
This is a linguistic puzzle involving
grammar changes in [LANGUAGE]. You are
given examples which include a source
sentence, a grammar change to apply to the
source sentence, and a target sentence.
Your task is to generate the target
sentence for the final example.

Example 1:
Source: [SOURCE SENTENCE]
Grammar Change: [CHANGE TAGS]
Target: [TARGET SENTENCE]

(...)

Now generate the target sentence for this
example:
Source: [SOURCE SENTENCE]
Grammar Change: [CHANGE TAGS]
Target:
```

Table 1: Our base prompt that we use for experimentation. `[LANGUAGE]` is replaced with Bribri, Yucatec Maya, Guaraní, or Western Sierra Puebla Nahuatl.

Additionally, we conduct these experiments with two LLMs: GPT-4o (OpenAI et al., 2024b) and DeepSeek-v3 (DeepSeek-AI et al., 2025), and set the temperature to 0.

Reference Book In one experiment, we include the line "*You are also given additional information about the morphology and syntax of the language.*" and copied the 'Morphology and Syntax' sections for Bribri, Maya, and Guaraní from a reference book (Campbell, 2000). We did not include morphological and syntactic information for Nahuatl as the reference book documented Classical Nahuatl rather than Western Sierra Puebla Nahuatl. We test this addition to the prompt with 10 examples from the training data included. This experiment is partly inspired by MTOB, a benchmark on low resource machine translation for LLMs using a human-readable grammar book (Tanzer et al., 2024). In contrast to MTOB, the grammar descriptions we include are only a few pages long.

Part of Speech Tags We experiment with additionally including a part-of-speech tagged source sentences in our prompt, alongside the original source sentences, for Maya and Guaraní data. We utilize open source part-of-speech taggers released by Apertium to generate our part-of-speech tagged

data (Forcada and Tyers, 2016; Kuznetsova and Tyers, 2021; Pugh et al., 2023).

3.2 Chain of Thought

We also experimented with chain of thought (CoT) prompting (Wei et al., 2022), instructing the LLM to offer a step-by-step analysis to arrive at a solution. We tested CoT prompting using DeepSeek-V3 first in a zero-shot setting, followed by experiments with few-shot settings using 10, 20, and 25 examples.

Our approach involved processing sentences by providing predefined steps using a structured CoT prompt. We varied the number of few-shot examples to evaluate their impact on model performance. The methodology followed these key steps:

1. **Understanding the Source Sentence** – The model was instructed to analyze the input sentence in the target language.
2. **Identifying the Required Change** – The model was guided to recognize and interpret the intended transformation.
3. **Retrieving Few-Shot Examples** – We experimented with different numbers of few-shot examples ($n = 0, 10, 20, 25$). For our CoT experiments, examples are selected based on the number of overlapping change tags with the test example, as described in Bui and von der Wense (2024).
4. **Applying the Transformation** – The model generated the modified sentence step-by-step, following CoT reasoning.
5. **Output Formatting** – The final prediction was in a standardized format (`PREDICTED TARGET:`), which helped us with the extraction of results.

3.3 Ensembling

We create an ensemble system by utilizing majority voting to combine up to six LLM outputs. We decide on the specific configuration of LLM systems to include for each language by comparing the scores on the dev set of ensembling every combination of up to six of our LLM experiment outputs, including all our prompt configuration experiments and our CoT experiment using DeepSeek-V3 and 25 examples. We also compare sentence-level, token-level, and character-level majority voting strategies.

3.4 Pointer-Generator Network

As a contrastive system, we train a character-level pointer-generator LSTM utilizing a language tag and change tags as features (Bahdanau et al., 2016; See et al., 2017; Vinyals et al., 2015). Our pointer-generator network has 1 encoder layer, 1 decoder layer, an embedding size of 128, and a hidden layer size of 512. We train on all data including our additional Bribri data, and use a learning rate of 1e-3, dropout set to 0.3, and optimize with Adam (Kingma and Ba, 2017). Training is conducted with early stopping, and we evaluate using a model checkpoint saved after training for 31 epochs.

4 Submitted Systems

We organize our submitted systems as follows:

System 1 A majority voting ensemble of up to six systems selected based on dev set performance for each language. For Bribri, this is a token-level majority voting ensemble of four LLM outputs. For both Maya and Guaraní, this is a whole sentence majority voting ensemble of six LLM outputs. For Nahuatl, the best single system outperformed any ensemble of multiple systems, so we include only a single non-ensembled system for Nahuatl.

System 2 The best prompt configuration for DeepSeek-v3 for each language, selected based on dev set performance.

System 3 GPT-4o using the same prompt configurations as System 2.

System 4 The best prompt configuration for GPT-4o for each language, selected based on dev set performance.

System 5 This system is CoT prompting of DeepSeek-v3 with 25 included examples.

System 6 This system is our pointer-generator LSTM.

5 Results and Discussion

We present our results on the test set for all six of our systems in Table 2. Our LLM ensemble system, System 1, performs the best of our submitted systems and is declared one of two winning systems on this year's task, achieving the highest scores in the task for Maya and Guaraní. Compared to last year's winning systems for Maya and Guaraní, our System 1 achieves an additional 10.00 percentage

System	Bribri			Maya			Guaraní			Nahuatl		
	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
1	22.71	45.68	71.63	**63.87**	**84.03**	**93.87**	**43.68**	**57.2**	**86.83**	3.33	12.2	52.75
2	20.21	42.5	71.99	59.35	82.32	92.95	38.19	50.28	85.41	3.33	12.2	52.75
3	20.21	44.51	72.21	56.77	80.59	91.77	38.74	55.47	86.17	1.67	11.66	49.27
4	18.75	45.09	71.42	60.00	81.94	92.94	40.93	54.89	86.02	1.67	12.5	49.67
5	15.83	40.02	70.59	59.03	80.48	92.39	41.21	55.04	86.21	2.5	12.84	55.31
6	5.42	20.67	49.65	9.68	46.71	67.19	6.32	4.79	46.28	0	0.62	27.73

Table 2: Test set evaluation results for our six submitted systems. Winning scores in the task are in **bold**.

points in accuracy for Maya and an 9.06 percentage points in accuracy for Guaraní (Chiruzzo et al., 2024). Compared to our highest scoring single LLM system submissions, our ensembling strategy also provides an increase in accuracy of 2.50 percentage points for Bribri, 3.87 percentage points for Maya, and 2.75 percentage points for Guaraní.

5.1 LLM Choice

Our single LLM systems, Systems 2, 3, and 4, exhibit a moderate amount of variation in score, though all still perform higher than last year's best systems for Maya and Guaraní. For a clearer understanding on how our selection of LLMs affects our performance on this task, we conduct an additional experiment on the dev set using our base prompt with 10 examples to compare our system performance when using GPT-4 (OpenAI et al., 2024a), specifically the gpt-4-0614 snapshot available through the OpenAI API[2]. We also compare performance when using two additional snapshots of GPT-4o: gpt-4o-2024-11-20 and gpt-4o-2024-05-13[3]. Our GPT-4o systems by default use gpt-4o-2024-08-06. We report the results of this experiment in Table 3. As seen in the table, using GPT-4 and different GPT-4o snapshots result in some variation in performance on the dev set compared to the LLMs used in our submitted systems, but this variation is only to a small extent. This could indicate that the specifics of our prompting technique and our method of selecting training examples play a more significant role in the higher performance of our single LLM systems, rather than simply our choice of LLMs.

[2] https://platform.openai.com/docs/models/gpt-4

[3] https://platform.openai.com/docs/models/gpt-4o

5.2 Prompting Configurations

We record the results of our experiments in varying LLM prompt configurations, which were referred to in selecting the components of our submitted systems, in Table 4. Notably, increasing the number of training examples included in the prompt did not strictly increase performance, and leveraging part-of-speech tags and reference book information also does not have a clear impact on performance as evaluated on the development set. Future work could take a fine-grained approach to understanding how such prompt configurations affect model predictions.

5.3 Nahuatl Performance and Future Work

We observe poor performance on Nahuatl across all of our experiments and submitted systems, compared to our performance on the other languages included in this task. One possible hypothesis as to why performance is so low is due to the extent of variation within Nahuatl and the extent to which LLMs have been trained on and can differentiate Nahuatl varieties. Western Sierra Puebla Nahuatl, the Nahuatl variety included in this task (de Gibert et al., 2025), is one of 30 varieties within the "language grouping" of Nahuatl recognized by the Instituto Nacional de Lenguas Indígenas (INALI). INALI further states that each language variety should be treated as languages themselves, particularly for educational matters, as well as in other areas including justice and health (INALI, 2008). Thus, in the spirit of this task, we propose that future work in developing systems to create educational materials for Indigenous language take a more variety-specific approach to Nahuatl, that may include sourcing and incorporating grammatical information about Western Sierra Puebla Nahuatl, and also possibly fine-tuning LLMs on Western Sierra Puebla Nahuatl data. Additionally, to understand the extent to which our systems are im-

Model	Bribri			Maya			Guaraní			Nahuatl		
	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
DeepSeek-V3	18.40	45.79	66.69	55.70	77.31	91.31	41.77	52.70	86.41	2.27	8.13	42.58
gpt-4o-2024-08-06	18.40	45.96	65.13	54.36	76.39	90.41	39.24	49.24	84.70	1.14	6.38	38.88
gpt-4o -2024-11-20	16.51	44.73	65.42	55.70	77.11	90.84	45.57	51.89	86.31	3.41	6.18	40.19
gpt-4o-2024-05-13	17.45	46.10	65.75	57.05	78.48	91.02	39.24	49.24	85.45	1.14	5.75	39.72
gpt-4-0613	18.40	46.25	66.54	56.38	76.20	90.93	37.97	50.68	83.49	2.84	5.71	38.95

Table 3: Results on the dev set of our comparison experiment with GPT-4, 3 different GPT-4o snapshots, and DeepSeek-V3, using our base prompt and 10 examples. Our submitted systems use the gpt-4o-2024-08-06 snapshot and DeepSeek-V3.

pacted by the linguistic diversity within Nahuatl, future analysis could examine whether incorrect outputs of our LLM-based systems are valid for other Nahuatl varieties. Such analysis may provide insight into how systems can be modified to better support Western Sierra Puebla Nahuatl specifically.

6 Conclusion

We presented the results of JHU's submission to the 2025 AmericasNLP shared task on the creation of educational materials for Indigenous languages. In developing our systems, we conducted experiments using different prompting configurations with GPT-4o and DeepSeek-V3, combined chain of thought prompting techniques with few-shot prompting, trained a pointer-generator LSTM, and construct a majority voting ensemble of LLMs. We achieve the highest performance on Maya and Guaraní with our ensemble system, which is declared one of two winning systems on this year's task.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Minh Duc Bui and Katharina von der Wense. 2024. JGU mainz's submission to the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 195–200, Mexico City, Mexico. Association for Computational Linguistics.

George L. Campbell. 2000. *Compendium of the World's Languages, Second Edition*, volume 1. Routledge, 29 West 35th Street, New York, NY 10001.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Ona de Gibert, Raul Vazquez, Robert Pugh, Abteen Ebrahimi, Pavel Denisov, Ali Marashian, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages. In *Proceedings of the 5th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2025)*, Albuquerque, New Mexico. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

Mikel L. Forcada and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Coleman Haley. 2024. The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure? In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 174–178, Mexico City, Mexico. Association for Computational Linguistics.

INALI. 2008. Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de méxico con sus au-

Prompt Config	Model	Bribri			Maya			Guaraní			Nahuatl		
		Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
3 examples	GPT-4o	16.98	44.52	62.89	54.36	77.37	90.84	39.24	47.44	85.39	**1.14**	4.86	34.98
	DeepSeek-V3	15.09	41.84	63.71	57.05	79.04	91.36	39.24	49.32	85.70	2.84	7.34	40.68
5 examples	GPT-4o	17.92	44.49	63.87	55.03	76.67	90.43	41.77	49.75	85.97	**1.14**	5.41	37.52
	DeepSeek-V3	17.92	46.24	65.75	55.03	76.96	90.61	**44.30**⋆	52.88	87.01	2.27	6.03	40.84
10 examples	GPT-4o	**18.40**	**45.96**	65.13	54.36	76.39	90.41	39.24	49.24	84.70	**1.14**	**6.38**	38.88
	DeepSeek-V3	18.40	45.79	66.69	55.70	77.31	91.31	41.77	52.70	86.41	2.27	8.13	42.58
20 examples	GPT-4o	15.57	44.76	64.75	**58.39**	**78.64**⋆	**90.98**	40.51	54.51	86.17	**1.14**	5.71	**39.19**
	DeepSeek-V3	**18.87**⋆	**47.68**⋆	**67.43**⋆	56.38	78.20	91.49	**44.30**⋆	54.02	**87.42**⋆	**5.11**⋆	**8.88**⋆	**43.56**⋆
3 ex. + POS	GPT-4o	-	-	-	53.02	76.24	89.16	39.24	**56.78**⋆	85.40	-	-	-
	DeepSeek-V3	-	-	-	55.03	77.29	90.59	36.71	49.25	83.63	-	-	-
5 ex. + POS	GPT-4o	-	-	-	48.32	72.51	88.75	41.77	55.96	85.12	-	-	-
	DeepSeek-V3	-	-	-	55.70	77.23	90.47	37.97	50.40	85.90	-	-	-
10 ex. + POS	GPT-4o	-	-	-	55.70	76.49	90.44	**44.30**⋆	52.37	86.15	-	-	-
	DeepSeek-V3	-	-	-	55.70	77.41	91.17	41.77	51.59	86.45	-	-	-
20 ex. + POS	GPT-4o	-	-	-	56.38	77.24	90.36	41.77	51.77	86.27	-	-	-
	DeepSeek-V3	-	-	-	**59.06**⋆	**78.55**	**91.54**⋆	40.51	51.08	86.21	-	-	-
10 ex. + book	GPT-4o	16.98	45.63	**65.86**	54.36	76.92	90.79	43.04	55.15	**86.95**	-	-	-
	DeepSeek-V3	16.04	44.47	65.90	55.03	76.50	90.95	43.04	**56.26**	86.33	-	-	-

Table 4: Results from experimenting with different prompt configurations using GPT-4o and DeepSeek-V3. The highest scores for each model on each language are in **bold**. The best scores across both systems for each language are indicated with a ⋆.

todenominaciones y referencias geoestadísticas. Instituto Nacional de Lenguas Indígenas, México, D.F.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Anastasia Kuznetsova and Francis Tyers. 2021. A finite-state morphological analyser for Paraguayan Guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89, Online. Association for Computational Linguistics.

Carla Victoria Jara Murillo. 2018. *Gramática de la lengua bribri*. EDigital, San José. Reviewed in *Revista de Filología y Lingüística de la Universidad de Costa Rica*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024b. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Robert Pugh, Francis Tyers, and Quetzil Castañeda. 2023. Developing finite-state language technology for Maya. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 30–39, Toronto, Canada. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic

expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (Americas-NLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.