Multilingual Emergency Alerts Translation and Accessibility

Natasha Schimka
University of Washington
nschim2@uw.edu

Jen Wilson
University of Washington
jenwils@uw.edu

Abstract

Emergency alerts in the United States are an essential means of delivering important information in the case of a local or national emergency. These messages are sent in English, which means that millions of people are not receiving alerts in languages they can understand, putting their life and health at significant risk. In this paper, we examine the language and associated risk gaps with current emergency alert systems and evaluate solutions to improve reach. We assessed machine-generated translations of emergency messages using BLEU, COMET, BERTScore, and ROUGE to compare the performance of DeepSeek, Gemini, ChatGPT, and Google Translate. We find that overall Google Translate has the most reliable performance.

1 Introduction

Emergency alerts are a critical tool used in the United States to send out important information in case of a local or national emergency. These are the messages we receive on our cell phones, hear on our radios, or see on our televisions. They convey information on an emergency situation that could have immediate impact on our safety, health, or life.

In 2024, the Pew Research Center (2024), reported that approximately 98% of people in the United States own cell phones. With this amount of coverage, it is possible to reach nearly every person in the country with an emergency alert when the need arises.

To amend that statement, it is possible to reach nearly every person in the country with an emergency alert in **English** when the need arises. According to the US Census (2020), over 27 million people reside in the United States and speak English "less than well." Despite significant advances in language technologies, millions of people are not receiving alerts in their languages. In this project, we analyze the language gaps and evaluate possible solutions to fill that gap.

We are interested in these research questions:

1. Which languages are most in need of translation within the state of Washington?
2. Can free tools such as LLMs and Google Translate provide accurate translations in an emergency?

2 Emergency Alerts

In this section we review the process for disseminating national and local emergency alerts.

2.1 Current Status

The Integrated Public Alert and Warning System (IPAWS) is an opt-out system that supports alerting the public in four different ways (FEMA, 2025):

- **Emergency Alert System (EAS)**: The message dissemination pathway that broadcasts alerts via cable, TV and radio.
- **Wireless Emergency Alert (WEA)**: The message dissemination pathway that broadcasts alerts and warnings to cell phones and other mobile devices. Most phones ship with the ability to receive these messages pre-installed.
- **NOAA Weather Radio (NWR)**: The message dissemination pathway that broadcast alerts and warnings via Non-Weather Emergency Messages (NWEM).
- **Internet-Based Services**: IPAWS also supports distributing alerts and warnings through internet-based services and existing alert systems such as sirens, wall beacons, desktop alerting, mobile applications and digital signs.

The primary goal of IPAWS is to transmit messages in emergency situations with information that could save life and health. A diagram of the IPAWS process is seen in Figure 1.

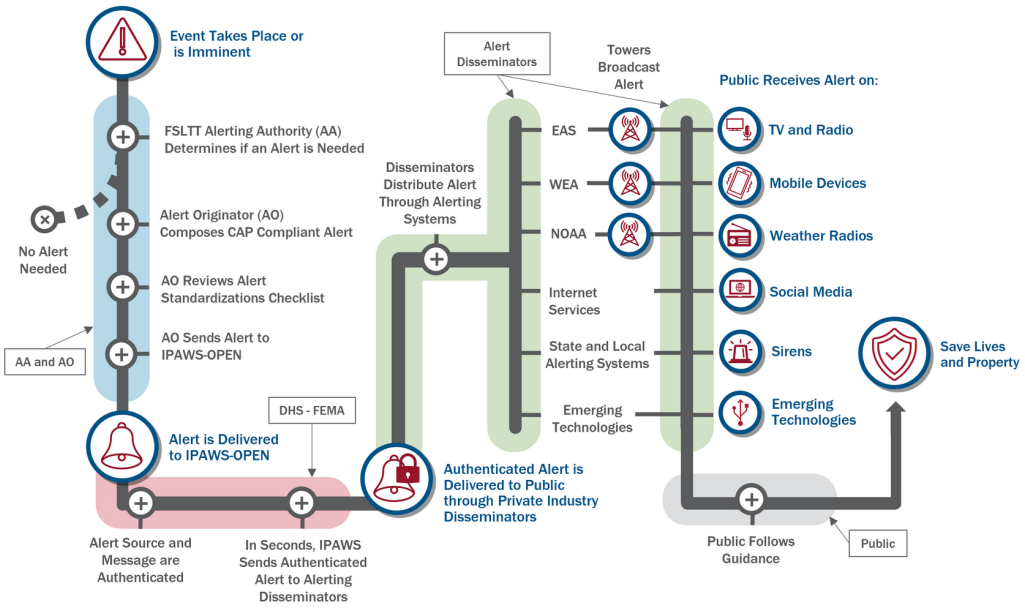

Figure 1: IPAWS alert distribution process (FEMA, 2025)

2.2 Local Mass Notification Systems

Some jurisdictions provide their own public alerting system. This allows agencies to control the type, language, and timing of the message. Examples of these systems near Seattle, Washington, include Alert King County, AlertSeattle, and UW Alerts.

These systems allow participants to register for alerts in the language of their choice, but the systems are opt-in and participation has been limited. While there are approximately 2.3 million people living in King County, Washington and as of May 2025, the Public Information Officer for King County states that enrollment has been limited with only 80,000 people signed up (Badger, 2025).

Opting in requires prior knowledge of the service and time to sign up and participate. In times of emergency, people who do not have access to these services, or are too stressed to access them, will be uninformed about what they should do to stay safe.

2.3 Future Plans

Since 2023, the Federal Communications Commission (FCC) has been preparing a report on implementation requirements for multilingual WEA templates for the 13 most commonly spoken languages in the United States (Federal Communications Commission, 2025a). After publication, participating wireless providers would have 30 months to support the new multilingual alerts. Providers would then need to install all translated templates on customers' devices, wherein the appropriate template is displayed based on the default language of the phone. (Federal Communications Commission, 2025a). [1]

While it is important that these templates be implemented to improve outreach, the proposed FCC report does not fully meet the needs of individuals. The Accessible Alert & Warning Workshops Report (Regional Disaster Preparedness Organization, 2024) notes several issues within the proposed FCC templates raised by participants, including words lacking direct translations and gaps between generational understanding of language and translation. In the following section, we'll further discuss the languages spoken by individuals in Washington State and the translation gaps with emergency alerts.

3 Translation for Emergencies

3.1 Translation Needs

The non-English languages covered by the FCC templates are: Arabic, Chinese (Simplified and Traditional), French, German, Haitian Creole, Hindi, Italian, Korean, Portuguese, Russian, Spanish, Tagalog, and Vietnamese. These languages were chosen because they are the top thirteen non-English languages spoken in the country.

We collected data to determine the approximate number of Low English Proficiency (LEP) peo-

[1] As of May 2025, the report has not yet been published.

ple in Washington State. We used data from the 2015 American Community Survey, because it is the most recent dataset that includes language proficiency to this level of detail, and analyzed the languages spoken by people living in Washington State counties who speak a non-English language and also speak English "less than well." The distribution of people and languages across counties is shown in Figure 2 and Appendix D. These figures show that while some of the FCC's languages are indeed highly needed by LEP individuals, such as Spanish, Chinese, and Tagalog, others like French and German have a much smaller need. Some languages, like Japanese, have a high need but are not available as translated templates.

According to the US Census Bureau (2023), there are now approximately 650,000 LEP individuals in Washington State.

3.2 Natural Hazards

We used data from the National Risk Index compiled by FEMA (2023) to determine the risk type and level for each of the 39 counties in Washington State. We simplified this list by categorizing any risk that was above zero as positive, with all othersbeingare negative.

We then examined the hazard risk for each county and compared it with the natural hazards templates in the FCC report. This allowed us to develop a list of natural hazards for each county and determine if there was a FCC template available for the hazard. For instance, King County is at risk of heat waves, ice storms, and landslides, none of which are available as FCC templates. This information is available in Appendix D.

3.3 Inadequate Alert Coverage

By combining language data and natural hazards data, we demonstrate that over half a million individuals in Washington State are at risk and need to have an alert in a language other than English. However, there is no system in place that can serve these individuals emergency information in an accessible language. Even after the FCC multilingual templates are published and implemented, there are other disasters specific to locales, such as drought or volcanic activity in Washington state, as well as other languages that are still in need. Outside of the IPAWS system, local alternatives are opt-in and significantly less capable at reaching a broad number of people.

4 Methods/Experiments

In this section, we detail our model design, prompt engineering, and data processing steps.[2]

4.1 Model and Parameter Selection

We hypothesized that if the need to communicate about a disaster in an unfamiliar language arose, officials might turn to a tool like ChatGPT, a web-accessible Large Language Model (LLM) that can generate text in response to a prompt passed as context. This suspicion was affirmed by our interview with the Public Information Officer at King County Office of Emergency Management, who mentioned that agencies might post an alert on social media, like Facebook or X, and use that website's embedded translation to translate the message (Badger, 2025).

We selected chat-based LLMs based on their ease of accessibility for emergency management officials, our target user, and out of those chose ones that offered some level of free API access: ChatGPT, DeepSeek, and Gemini. Lastly, we chose Google Translate as a non-LLM method for comparison, as it's similarly accessible and uses Neural Machine Translation (NMT) (Wu et al., 2016). Details on how we accessed and configured these systems are available in Appendix A.

We evaluated five disasters that would affect Washington state—fire, extreme wind, flood, a boil water notice, and a 911 outage. We chose these situations because they reflect a wide variety of alerts, simple and complex alike. The 911 outage alert is short and informs people that 911 is unavailable; the boil water notice is more complex because it includes methods on how to make water safe to drink.

Per each language-disaster-prompt combination we collected five responses, with the exception of Google Cloud Translation, which applies translation directly without the use of prompts. There, we collected five responses per language-disaster pair, for a total of 2330 responses out of the 2375 we intended to capture.[3]

4.2 LLM Prompt Engineering

Prompt engineering, or customizing the prompt submitted to an LLM, has a large impact on an LLM's ability to provide concise and relevant information (White et al., 2023). For this reason,

[2] https://github.com/zepam/MultiLingualEmergencyAlerts

[3] DeepSeek data was incomplete due to rate limits.

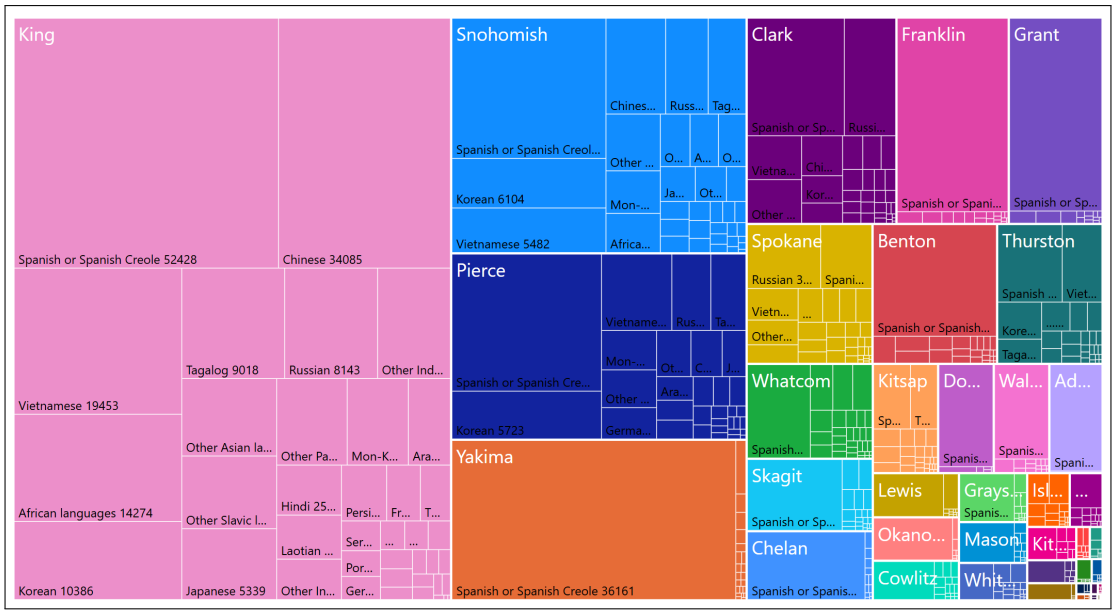

Figure 2: Low english speakers by Washington State county (U.S. Census Bureau, 2015)

we wrote six different prompts to evaluate whether our selected strategies would result in an improved translation. These prompts are included in Appendix B.

1. **Prompt A (simple)**: the simplest possible prompt that a user might submit to an LLM. This is our naive comparison.

2. **Prompt B (persona)**: the persona pattern instructs the LLM to generate its output from the perspective of a certain persona to help constrain the generated text (White et al., 2023). We asked the LLM to imagine it worked at the FCC.

3. **Prompt C (one-shot)**: providing a template can guide the LLM's output in terms of style and content (White et al., 2023). We included a direct example of an FCC alert template before asking the LLM to generate one.

4. **Prompt D (chain of translation)**: chain of translation is a technique that can improve the translation of low-resource languages when performing a task such as annotation or classification. The prompt is modified to ask the LLM to translate the input context to English first, perform the task on the English translation, and provide the output in the original language (Deshpande et al., 2024). Our own task can be thought of as two steps—generating a prompt and translating it, and so we ask the LLM to generate the prompt in English first, then translate it to the target language.

5. **Prompt E (cross-lingual alignment)**: chain of thought (CoT) is a prompting technique that instructs the LLM to think about a given task step-by-step and can improve performance of single-language tasks. Cross-lingual alignment is a zero-shot CoT technique developed specifically for multilingual contexts through two components: ask the model to understand the task in English step by step as an expert speaker of the target language (Qin et al., 2023).

6. **Prompt F (direct translation)**: Given the FCC template for a particular disaster, we ask the LLM to translate it directly to the target language.

Throughout the rest of this paper, we will refer to prompts A-E as "non-translate prompts" and prompt F as "direct translation prompt."

4.3 Data Processing

Although Arabic is read from right-to-left, the LLMs nearly always printed it out from left to right judging by the placement of punctuation along the right-hand side of words. We applied the Python `arabic_reshaper` library to ensure the text was not broken and then used the `get_display` method from `bidi` to reverse the text.

Some output included irrelevant characters, such as emoji and URLs to unrelated Github repositories. That is not information an alert would typically contain. Despite the likelihood of these characters lowering the scores, we chose to leave them in the evaluation data because they are a nuisance that emergency officials would need to clean up themselves.

We instructed LLMs not to translate the fillable variables, such as [LOCATION], but in the event that this directive was not followed, we left these variables in place. As we lacked time to fully clean up the response data in a script, the only characters that were removed were newlines `\n`, and for Google Translate, we did strip out fillable variables because Google Translate can't be configured to leave them in English.

For DeepSeek particularly, the quality in responses varied the most in quality. Sometimes it would output a sequence of 1s, text in English (or Chinese), or the exact same Github repository. For this reason, we manually removed these English characters from the data as we saw them, but the presence of this data could negatively impact the scoring of its generated translations.

5 Evaluation

In this section we review our methods of evaluation for both our human evaluation pipeline and our metric evaluations. All automated metrics were implemented using Huggingface's `evaluate` library.[4] Tokenizers from `sacrebleu`[5] were applied to ROUGE and BLEU calculations—`TokenizerZh` for Chinese, `Flores101Tokenizer` for Arabic and Vietnamese, and `TokenizerV14International` for Spanish and Haitian Creole.[6]

5.1 Human Evaluation on References

We had two possible sets of templates to use as gold standards: the FCC templates, prepared with the assistance of native speakers and professional translators (Federal Communications Commission, 2025b), and an extensive list of translated alerts created by the King County Office of Emergency Management (2020). As the authors do not speak the languages included in the FCC templates, with the exception of Spanish, we had each translated alert evaluated by humans to sanity-check the quality of these potential gold standards.

We selected five languages: Haitian Creole, Arabic, Vietnamese, Chinese (Traditional), and Spanish. For each of these languages, we had access to native speakers; the Chinese speaker specifically spoke Mandarin.

For Spanish, Vietnamese, and Chinese, we had alert templates prepared by both the FCC and King County and both sets of alerts were evaluated for these languages. The Arabic and Haitian Creole alerts were available only in the FCC report.

We adapted the Multidimensional Quality Metrics (MQM) framework for human evaluation. MQM was introduced in 2021 by Freitag et al. as a framework to provide a customizable list of translation errors that can be tailored to the specific occasion and scored. We selected a set of ten possible translation errors and categorized them as minor or major, as shown in Table 5. The Spanish evaluation was completed by one of this paper's authors and the others were done by coworkers or friends.

We used a simple MQM scoring system that considered only the total scores for each language and alert combination. The Vietnamese, Haitian Creole, Spanish and Chinese alerts were all considered acceptable gold standards based on the evaluator scores.

Vietnamese was scored by two speakers who both saw no issues with the translations (0 points). Haitian Creole was scored by one person who said the flood alert was awkward (1 point). Spanish was scored by one person who noted four minor errors (4 points).

The Arabic evaluator rated the major issues with the flood alert (13 points) and extreme wind alert (12 points) and two minor issues with the Wildfire/Fire alert (2 points). These alerts were the only references available and we decided to use them despite these high scores.

Results from our six human evaluators are given in Appendix C. Because the FCC and King County translations were similar in quality, we chose to use only the FCC templates as our gold standard comparisons for consistency.

For Chinese, we accidentally included a mix of traditional (from the FCC) and simplified characters (from King County) in our human evaluation but ultimately chose traditional characters to be our gold standard based on a recommendation by the City of Seattle that traditional characters can serve

[4] https://huggingface.co/docs/evaluate/en/index

[5] https://github.com/mjpost/sacrebleu

[6] The selected tokenizer impacted ROUGE and BLEU scores a fair amount, which we'll revisit in the discussion section.

both Mandarin- and Cantonese-speaking populations (City of Seattle, 2020).

5.2 BLEU

BLEU is traditional MT metric that compares a gold standard translation (a reference) with a candidate sentence and assigns a numerical score between 0 and 100 based on *n*-gram overlap and a brevity penalty that penalizes candidates that are shorter than the reference (Papineni et al., 2002). While BLEU is still commonly applied due to its ease of usage and fast calculation, it will penalize paraphrases and other such still-fluent translations and has lead to a desire for more modern evaluation metrics (Rei et al., 2020). In terms of interpretation, 40 and above is considered to be a high quality translation and 60 and above is human-quality or better (Google, 2025).

5.3 COMET

The COMET scoring method has been proposed as a modern alternative to BLEU. It leverages pretrained embeddings from a multilingual XLM-RoBERTa model to evaluate the semantic similarity between the reference and target translation (Rei et al., 2020). Because all languages in this project are included in the model, we are able to calculate COMET scores for each target language. COMET scales the scores to between 0 and 1. Scores close close to 1 indicate a high-quality translation and scores close to 0 indicate a translation that is similar to random chance.

5.4 BERTScore

Similarly to COMET, BERTscore uses contextual embeddings to calculate a cosine similarity between reference and target translations between -1 and 1. Baseline scaling is available for select languages, of which ours are included, to normalize the result to be between 0 and 1 (Zhang et al., 2020).

5.5 ROUGE

ROUGE is an n-gram based metric that is typically used for evaluating summarizations. Prioritizing recall, ROUGE-1 counts unigrams, ROUGE-2 counts bigrams, and ROUGE-L counts the longest common subsequence between the comparison texts (Lin, 2004). We chose to include ROUGE evaluations because our alerts are quite short and are templematic in nature; we wanted to penalize any translations that stray from the source material.

6 Results and Discussion

6.1 Prompt performance

In Figure 3, we break down our performance on BERTScore, ROUGE, and COMET on the non-translate prompts, averaged across all services excluding Google Translate. Our naive baseline, the simple prompt, had the lowest performance, particularly in ROUGE-2, but out of the remaining prompts, there appears to be no clear winner. Across all these prompts, both BERTScore and COMET were high, around 0.8 for the former and 0.7 for the latter, indicating that these translations are overall capturing the meaning of the original templates.

Figure 4 shows results for BLEU on these same prompts, this time broken down by each service. No language even approaches the "good" score of 40, suggesting that these translations are of lower quality. Scores for Chinese (traditional) are particularly low for all three services. We will return to these results in the discussion section.

Interestingly, the results change significantly when we look specifically at the direction translation prompts, including Google Translate, in Figure 5. Whereas ROUGE scores never exceeded 0.5 in the earlier set of data, these prompts consistently are in the 0.7 range for all disaster types. BERTScore in particular is close to 1 on the wind, flood, and boil water notice alerts.

This trend continues when we focus in on the BLEU scores for the direct translation prompts in Figure 6. Whereas the other prompt engineering techniques never approached a score of 40, many are now even approaching 60, indicating a human-quality translation. That said, there are still some disasters that fall behind in performance, such as extreme wind for all services. Out of all four services, Google Translate most frequently scored under 40, particularly for Spanish and Vietnamese.

6.2 Service performance

Lastly, we break down our results by language and service in Figure 7 for all metrics except BLEU. When averaged across all disasters and prompts, no single service performed demonstrably worse on any particular language, and the performance of ChatGPT, DeepSeek, and Gemini was relatively the same on all languages. Although those same services performed better than Google Translate on the direct translation prompts, their lower performance on all other prompts ultimately lead to lower

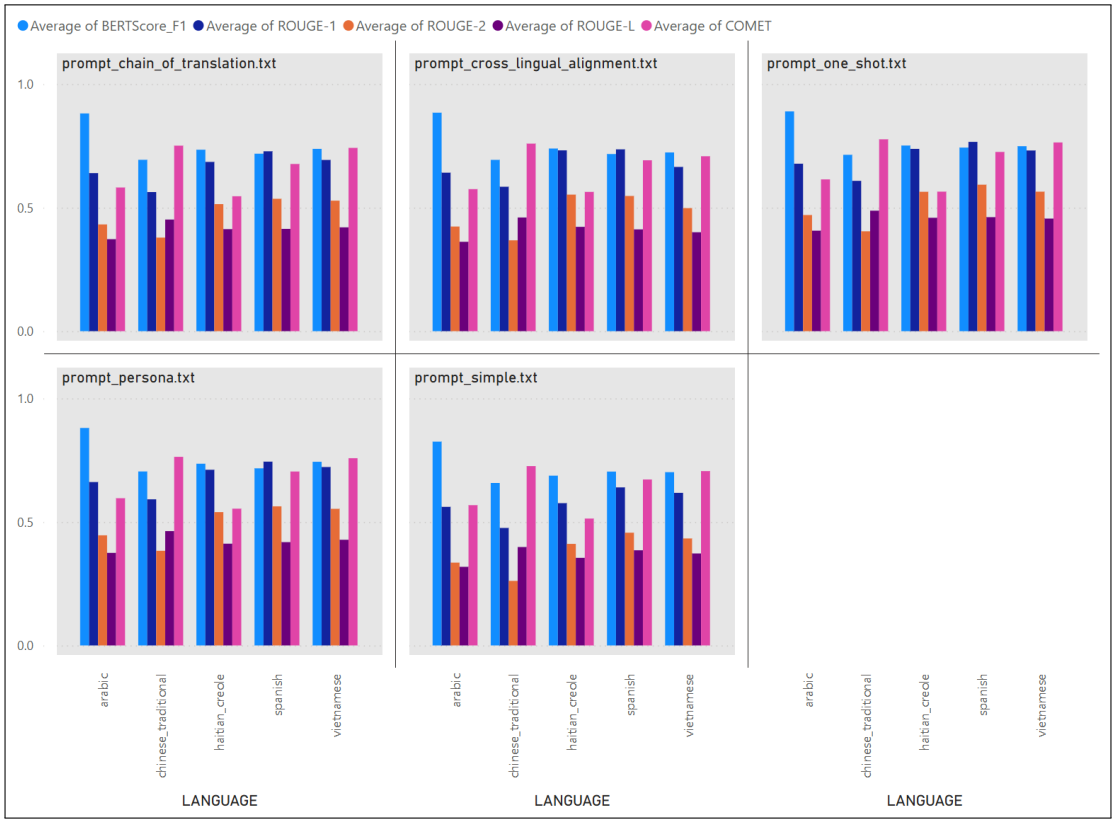

Figure 3: Relevance scoring on the non-translate prompts, broken down by language and averaged across disasters and services. Languages are indicated along the bottom. Each section indicates performance of a particular prompt for all languages.

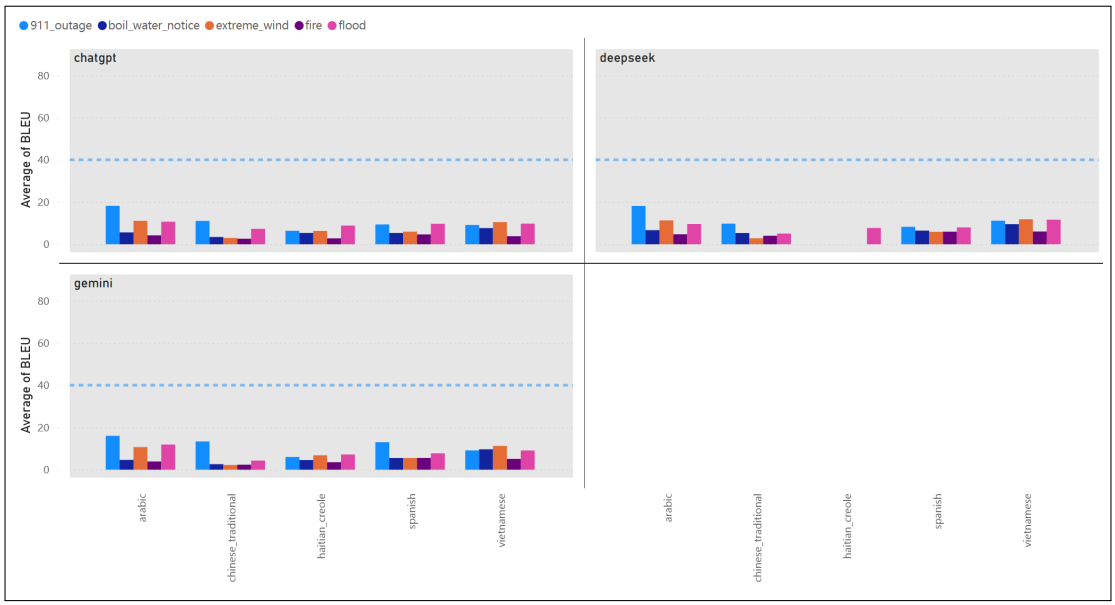

Figure 4: BLEU scores on the non-translate prompts, broken down by disaster, service, and language. Data for DeepSeek - Haitian Creole is incomplete.

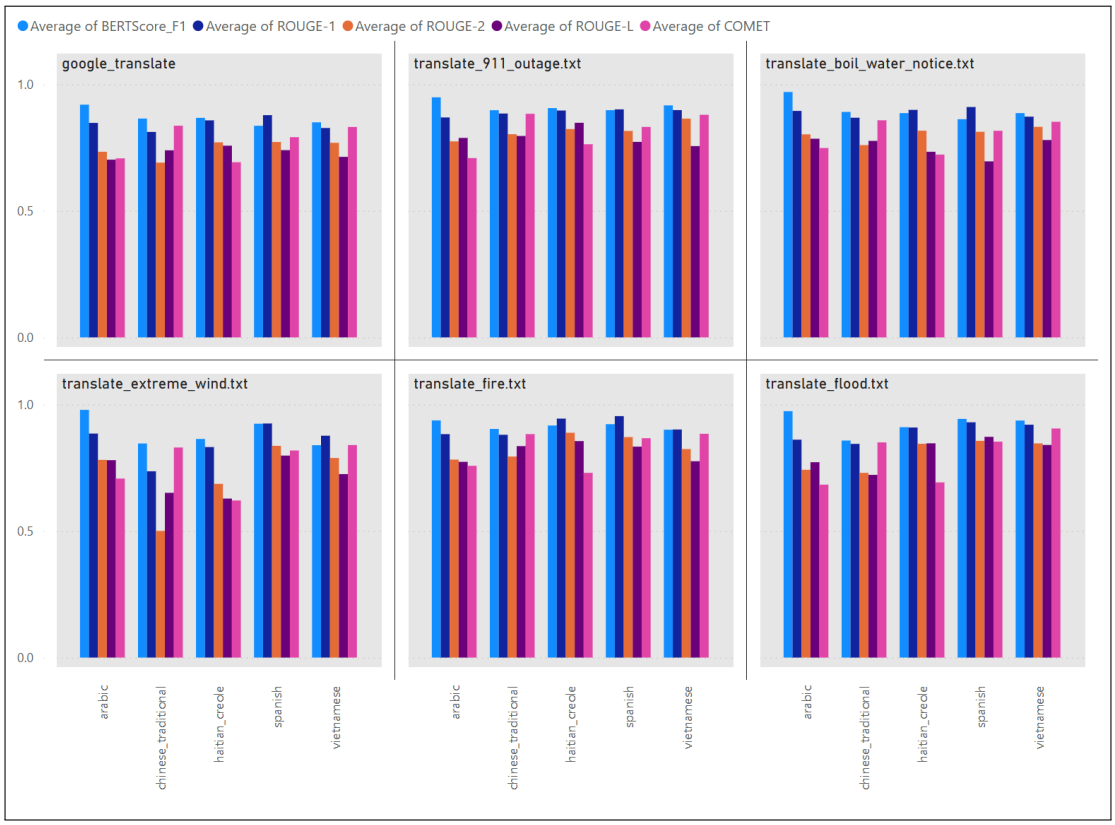

Figure 5: Relevance scores on the direct translation prompts, broken down by language and disaster and averaged across services.

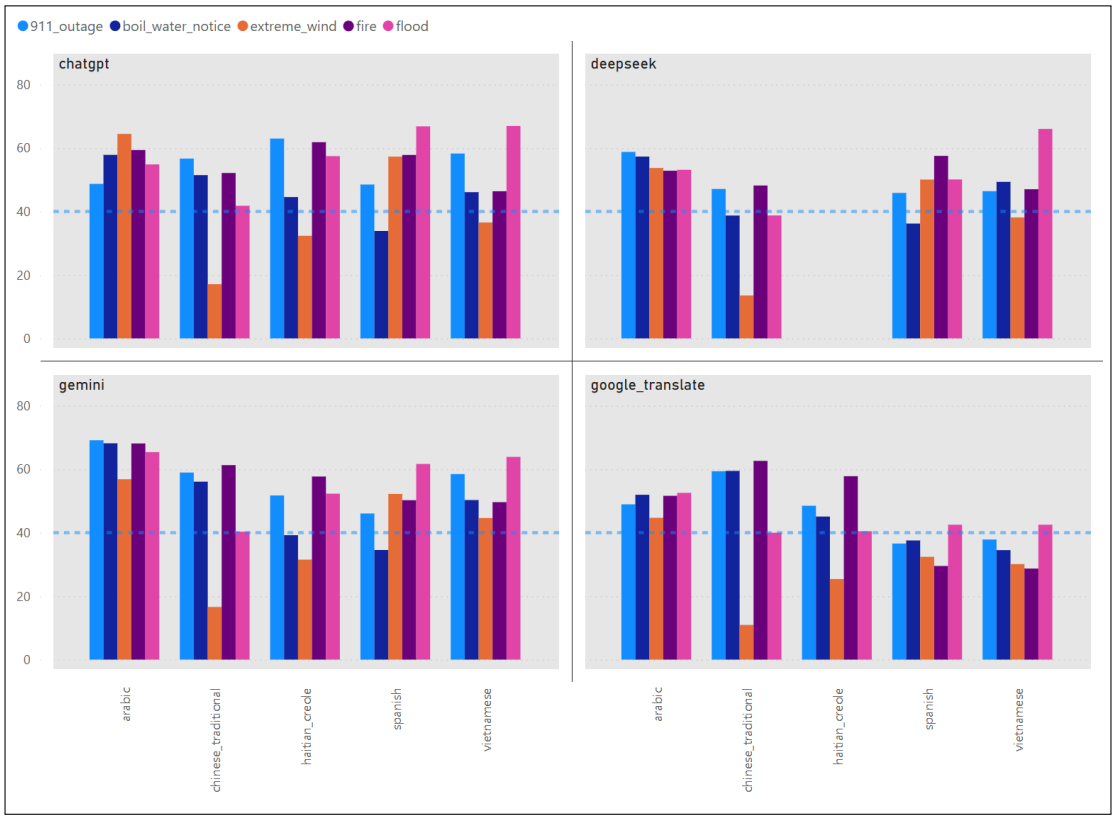

Figure 6: BLEU scores on the direct translation prompts, broken down by disaster, service, and language. Data for DeepSeek - Haitian Creole is incomplete. The blue line at 40 indicates a good score.

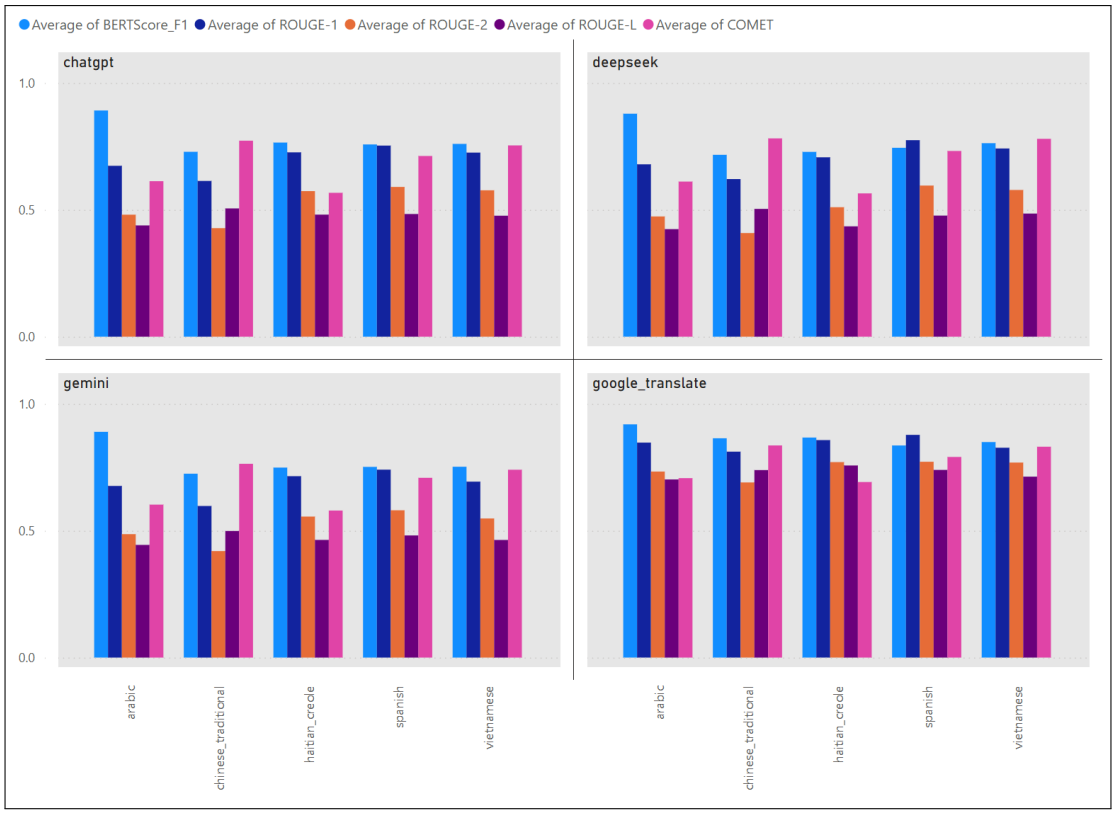

Figure 7: Comparison of the overall performance of each translation service on our task.

overall scores compared to Google Translate.

6.3 Discussion

The drastic difference in BLEU scores between the translate and non-translate prompts merits a deeper look. We compared the average length of translated Chinese responses from ChatGPT to our gold standards for the same disaster in Table 1. We found that ChatGPT's translations were always less than 100, which were in turn always shorter than the target reference. Because BLEU applies a brevity penalty for translations that are shorter than the target sentences, this negatively impacted our BLEU scores. In retrospect, we explicitly instructed our LLMs to generate alerts that were less than 100 characters in length but neglected to ensure that this character limit matched the longest alert in our list of gold standards. By contrast, the direct translation prompts did not request a character limit and performed far better.

Overall, the most effective prompting strategy was to request translation of an existing FCC alert from English to the target language. The high scores across the board, particularly the BLEU scores in the 60+ range, seen in Figure 6, suggests that these alerts may exist in some form within the training data for ChatGPT, Gemini, and DeepSeek. While these particular templates from the FCC were only finalized January 2025, the format of these alerts is quite standardized and it may be that earlier variants were included in the training data.

In most cases, it is considered bad technique to have the test data (the translated alerts) in the training data. Ultimately, the goal for emergency alerts is to get critical information out to the people who need it. In this case, the goal is to get quality information out of the service, whether the alert is being translated or merely output from the service. If the alerts are in the training data, then the output would be consistent regardless of whether or not it is actually being translated. However, because the training data for these LLMs is not public, it then becomes difficult to predict whether a particular language and disaster pairing that has not been covered in this paper will be translated adequately. Due to this fact, even though ChatGPT does boast higher BLEU scores on Arabic for instance than Google Translate, Google Translate's performance across the board is relatively good and is our recommended service for emergency response officials based on the results in this paper.

Google Translate's sub-40 BLEU score on Span-

Disaster name	Gold standard length	Average generated length
Flood	107	93.97
Extreme wind	128	97.62
Fire	188	95.93
Boil water notice	187	99.43
911 outage	115	97.89

Table 1: Comparison of the average generated length for Chinese responses for ChatGPT compared to our gold standard from the FCC. Our non-translate prompts requested the alert response to be limited to 100 characters, which did not necessarily match the actual dataset.

ish was unexpected. We ran a small experiment using different `sacrebleu` tokenizers within the Bleu calculation and found that it did make a difference, even for languages that are broken by whitespace like Spanish and Haitian Creole. In Table 2, we show that using the FLORES-101 evaluation benchmark (Goyal et al., 2021) improves the BLEU scores for all languages except Chinese, particularly Arabic. We similarly needed to apply this tokenizer to ROUGE calculations to get non-zero ROUGE scores for Arabic. However, we encountered memory issues attempting to run our evaluation script using FLORES-101 for all languages but Chinese, and so our previous graphs instead use the tokenizers outlined in Section 5.

7 Recommendations

We offer three recommendations to address the translation gap with emergency alerts:

1. Continue to push the FCC to provide multilingual alerts. All efforts to get the FCC report published and to encourage wireless providers to install alert translations are worthwhile.

2. Use the [URL] field on an alert template to link to a page that provides the alert in English while also providing translations on the same page in all languages with templates. This would guarantee that the provided alerts all meet the translations standards of the agencies.

3. On the page with the translated alerts (noted in #2), provide a Google Translate drop-down to cover additional languages.

We acknowledge that the translations may have errors and there is a concern that the emergency information may be inadequate. Our results in this project indicate that it is likely the alerts will be translated correctly.

8 Future Work

The results of the output capabilities these services are promising and this work should be extended to all languages that meet the criteria in the 2017 Washington State law[7] that requires agencies develop plans to send emergency alerts in languages that make up 5% of the population or 1,000 residents, whichever is less. This list includes more than 27 languages in Washington State (King County Emergency Management, 2022).

The BLEU evaluation could be expanded to use multi-reference evaluation to take advantage of the templates from King County.

9 Conclusions

In this paper we demonstrate that translation quality from Google Translate is reliable enough that agencies can feel confident that alerts will be translated adequately. Even though we do not suggest that agencies use ChatGPT, Gemini, or DeepSeek for translations, the general public will get fairly good translations if the individuals choose to use those services.

Limitations

This paper was limited by the cost of LLM access; free student credits were used for Google and Microsoft Azure access. Due to these models being accessed via API, it may not exactly reflect the experience someone may have when issuing these same queries within the chat GUI due to differences in hyperparameters.

Additionally, we used the FCC multilingual templates as our gold standards for translation, but these templates ranged in quality and could have altered the results of the evaluation metrics.

With the exception of Vietnamese, our human evaluations were evaluated by only one person per

[7] https://app.leg.wa.gov/rcw/default.aspx?cite=38.52.070

Language	Alert Type	Tokenizer			
		flores101	intl	none	zh
Arabic	Average	**56.3408**	38.7938	33.2472	35.7884
	911_outage	**57.1998**	29.1957	22.7427	19.7265
	boil_water_notice	**54.3638**	36.4789	30.7257	35.6694
	extreme_wind	**54.4737**	38.9968	34.1958	37.5222
	fire	**54.3438**	38.5633	30.0244	35.2897
	flood	**61.3231**	50.7342	48.5475	50.7342
Chinese Traditional	Average	41.5990	17.8768	0.6221	**52.1092**
	911_outage	55.3517	35.5042	3.1105	**64.3047**
	boil_water_notice	57.2937	17.7877	0.0000	**64.0329**
	extreme_wind	3.6763	0.9134	0.0000	**13.5870**
	fire	56.9271	33.5770	0.0000	**70.5642**
	flood	34.7465	1.6018	0.0000	**48.0572**
Haitian Creole	Average	**57.0674**	51.9580	48.8788	52.9277
	911_outage	57.1693	58.4777	53.9148	**63.3262**
	boil_water_notice	**57.4459**	52.0941	49.9481	52.0941
	extreme_wind	**32.2415**	26.5642	22.4693	26.5642
	fire	**74.5956**	67.7610	65.5090	67.7610
	flood	**63.8846**	54.8930	52.5530	54.8930
Spanish	Average	**49.6372**	44.6411	39.8046	45.7699
	911_outage	42.6908	42.7374	35.6017	**49.2605**
	boil_water_notice	**48.0308**	42.9838	37.2965	43.2458
	extreme_wind	**41.7323**	38.1821	35.1680	37.0409
	fire	**45.0347**	38.3189	32.0972	38.3189
	flood	**70.6974**	60.9832	58.8598	60.9832
Vietnamese	Average	**44.8944**	42.6130	38.1356	43.2976
	911_outage	49.2931	49.4056	43.6175	**52.8286**
	boil_water_notice	**42.8122**	38.9118	33.6660	38.9118
	extreme_wind	**38.3456**	36.9139	29.9838	36.9139
	fire	**37.9546**	34.9381	29.4306	34.9381
	flood	**56.0664**	52.8957	53.9802	52.8957

Table 2: BLEU scores across five languages and emergency alert types, evaluated using four different tokenizers.

language. These speakers were not tested for language fluency.

The data to evaluate the languages with the highest translation needs were from a survey from 2015 and have likely changed since then. With possible federal budget cuts, datasets on this topic may not be available in the future.

Acknowledgements

We would like to thank our language evaluators who graciously shared their time and knowledge with us—Bernard, My, Ibrahim, and Celia—as well as Sheri Badger for sharing her time and answering all our questions about current needs for translation.

References

Sheri Badger. 2025. Public Information Officer, King County Office of Emergency Management. Teams meeting.

US Census Bureau. 2020. People That Speak English Less Than "Very Well" in the United States. (accessed 05-22-2025).

U.S. Census Bureau. 2023. Explore Census Data. (accessed 05-31-2025).

City of Seattle. 2020. Seattle top tier languages. (accessed 05-30-2025).

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages. *Preprint*, arXiv:2409.04512.

Federal Communications Commission. 2025a. Multilingual wireless emergency alerts. (accessed 04-14-2025).

Federal Communications Commission. 2025b. Wireless emergency alerts (wea). (accessed 05-13-2025).

FEMA. 2023. National risk index. (accessed 05-12-2025).

FEMA. 2025. Integrated Public Alert & Warning System. (Accessed 05-25-2025).

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021.

Experts, errors, and context: A large-scale study of human evaluation for machine translation. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1460–1474.

Google. 2025. Evaluate models. (accessed 06-03-2025).

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Preprint*, arXiv:2106.03193.

King County Emergency Management. 2022. King county regional inclusive emergency communications plan.

King County Office of Emergency Management. 2020. Inclusive emergency communications - King County, Washington. (accessed 05-07-2025).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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.

Pew Research Center. 2024. Mobile Fact Sheet: Tech Adoption Trends. Accessed 05-22-2025.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Preprint*, arXiv:2310.14799.

Regional Disaster Preparedness Organization. 2024. Accessible alert warning workshops report.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

U.S. Census Bureau. 2015. American community survey 5-year estimates subject tables, table b16001.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *Proceedings of the 30th Conference on Pattern Languages of Programs*, PLoP '23, USA. The Hillside Group.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, and 12 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Model Details

We accessed ChatGPT via the AzureOpenAI Python SDK and Gemini via Google's genai[8] SDK and the Gemini Developer API. While there is no direct API access to Google Translate, we used the Basic version of Google Cloud Translation[9] as an approximate equivalent. DeepSeek was accessed via OpenRouter[10], an SDK that can access many models through a unified interface and offers 50 free requests per day.

Because our intended users would access these interfaces via a GUI rather than the API, we aimed to mimic the hyperparameters of the chat GUI but encountered difficulties finding documentation publicizing these numbers. For each LLM call, we passed a temperature of 1.0, top p of 1.0, set the max tokens to 2048, and passed a role of "user" to the prompt. Where possible, we used the model version that's presented as the default option within the respective web interface for the LLM service as shown in Table 3.

B English Prompts for Multilingual Generation

Here we include the final prompts that were used to elicit responses from our translation services. To reduce the amount of extra information the LLM returned, each prompt ends with a request to only return the alert. Each prompt also includes placeholder text for DISASTER and LANGUAGE, which were replaced in-line before sending the prompt to the translation service.

Prompt A (simple):

[8] https://github.com/googleapis/python-genai

[9] https://cloud.google.com/translate/docs/reference/libraries/v2/python

[10] https://openrouter.ai/deepseek/deepseek-chat-v3-0324:free

Model name	Chosen model version	GUI model version
Google Cloud Translation	Basic	Google Translate
Gemini	2.0 Flash	2.5 Flash
DeepSeek	V3-0324	Likely V3
ChatGPT	2024-12-01-preview (GPT-4o)	GPT-4o

Table 3: Comparison between the model versions we chose and the default version available in the chat GUI.

```
Can you write an emergency text
about DISASTER in LANGUAGE
please?

Only return the text of the
alert in LANGUAGE. Do not
include any other response to
the prompt.
```

Prompt B (Persona):

```
Imagine you work at the FCC.
Write a short emergency alert,
no more than 100 words, about
DISASTER. Include fields for
[TIME], [LOCATION], and a [URL]
for more information so it can
be customized; these fields
should be in English.  The alert
doesn't need a header, just a
message body.

Only return the text of the
alert in LANGUAGE. Do not
include any other response to
the prompt.
```

Prompt C (one-shot):

```
Write a short emergency alert,
no more than 100 words, about
DISASTER. Include fields for
[TIME], [LOCATION], and a [URL]
for more information so it can
be customized; these fields
should be in English.  The alert
doesn't need a header, just a
message body.

Only return the text of the
alert in LANGUAGE. Do not
include any other response to
the prompt.

For example, here's how the
English template might look for
a tornado:

[SENDING AGENCY]: A TORNADO
EMERGENCY is in effect for
[LOCATION] until [TIME].
Tornado spotted in this area.
This is a life-threatening
situation.  Take shelter now
in a basement or an interior
room on the lowest floor of a
sturdy building.  If you are
outdoors, in a mobile home,
or in a vehicle, move to the
closest sturdy shelter and
protect yourself from flying
objects.  Check media.  [URL]
```

Prompt D (chain of translation):

```
Write a short emergency alert,
no more than 100 words, about
DISASTER. Include fields for
[TIME], [LOCATION], and a [URL]
for more information so it can
be customized; these fields
should be in English.  The alert
doesn't need a header, just
a message body.  Perform the
following steps:

1.  Generate the template in
English first 2.  Translate it
to LANGUAGE and output it

Only return the text of the
alert in LANGUAGE. Do not
include any other response
to the prompt, including the
template from step 1.
```

Prompt E (cross-lingual alignment):

```
Imagine you are an expert in
multilingual understanding
in LANGUAGE. Write a short
emergency alert, no more than
100 words, about DISASTER.
Include fields for [TIME],
[LOCATION], and a [URL] for
more information so it can
be customized; these fields
should be in English.  The
alert doesn't need a header,
just a message body.  Let's
understand the task step by step
in English!

Only return the text of the
alert in English.  Do not
include any other response to
the prompt.
```

Prompt F (direct translation):

There were five variations on this template, one for each disaster type. Here's an example:

```
Translate this emergency alert
to LANGUAGE. Don't respond
with anything outside of the
translation.  For the variables
in square brackets, leave those
in English.

[SENDING AGENCY]: An EXTREME
WIND WARNING is in effect for
[LOCATION] for the immediate
```

		Haitian Creole	Arabic	Spanish	Chinese (Traditional)	Vietnamese
Flood	FCC	1	13	1	0	0
	KC	n/a	n/a	0	1	0
Extreme Wind	FCC	0	12	0	0	0
Wildfire/Fire	FCC	0	2	1	0	0
	KC	n/a	n/a	1	0	0
Boil Water	FCC	0	0	1	0	0
	KC	n/a	n/a	0	0	0
911 Outage	FCC	0	0	0	0	0
	KC	n/a	n/a	0	2	0

Table 4: Total error score for each language and template source

```
danger of life-threatening winds
until [TIME]. Take cover NOW
in an interior room of a sturdy
building, away from windows.
Protect your head from flying
objects.  Do NOT go outside
if the wind calms!  Winds will
quickly become dangerous again.
[URL]
```

For Google Cloud Translation, which didn't require prompts, we supplied the English FCC templates for our set of targeted disasters.

C Human Evaluation Results

In Table 4 we include a summary of the errors reported by our volunteer evaluators. In the table, "A" refers to the FCC templates and "B" refers to the King County templates. For Chinese specifically, the "A" template was written with traditional Chinese characters and the "B" template was with simplified due to a mistake.

The most common type of error reported was an awkward style, indicating that these alerts would still be understood. Arabic was the only language that had major errors in the form of wrong terms, indicating that the alert may be misunderstood by the receiver.

D Additional Data

We built all graphics in PowerBI Desktop by Microsoft. If you have a UWNetID, you may log in and view the data here: https://uwnetid-my.sharepoint.com/:u:/g/personal/jenwils_uw_edu/EVo-q2fwthJJou3_5rgce8YBdCfqxdun2dY0PtQkTu5qqg?e=vzNmz4

Error type	Severity of error	Score
Wrong term	major	10
Undertranslation - content is inappropriately less specific than needed	major	10
Omission - information is missing in target	major	10
Unintelligible	major	10
Inconsistent use of terminology	minor	1
Overtranslation - content includes information not needed	minor	1
Grammatically incorrect	minor	1
Spelling errors	minor	1
Awkward style	minor	1
Wrong time or date format	minor	1

Table 5: Our rubric for assigning point values to each error.