LLM-Powered Automatic Translation and Urgency in Crisis Scenarios

⚠️ *“If a **crisis alert** is mistranslated, does it still save lives?”*

Belu Ticona
mticonao@gmu.edu

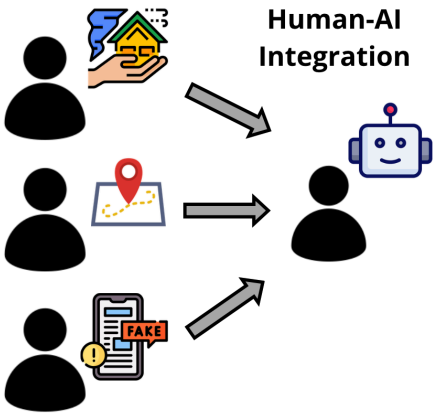
Antonios Anastasopoulos
antonis@gmu.edu

Speech+NLP+HCI

GEORGE MASON UNIVERSITY

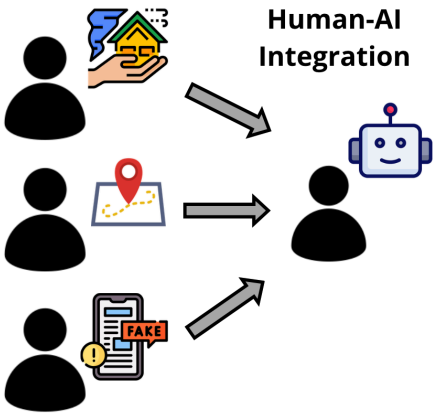
Why multilingual crisis NLP matters

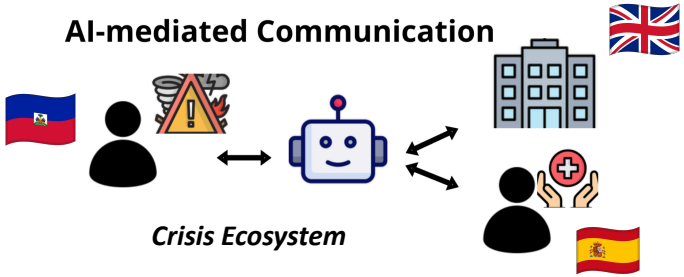
LLMs widely adopted as general-purpose systems



Why multilingual crisis NLP matters

LLMs widely adopted as general-purpose systems

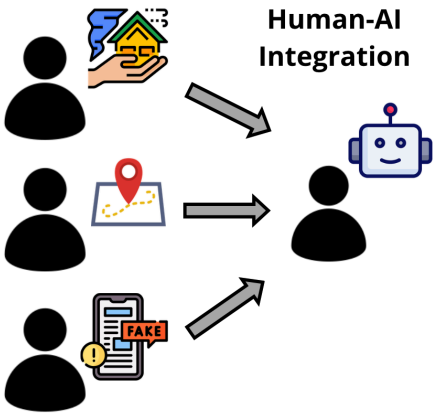




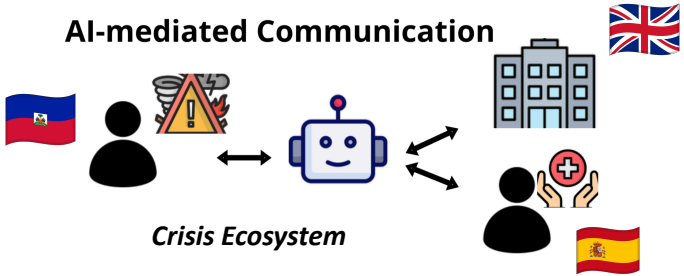
LLMs used differently by stakeholders

Why multilingual crisis NLP matters

LLMs widely adopted as general-purpose systems



English-only benchmarks



Exposure mismatch

LLMs used differently by stakeholders

Key Gap: No rigorous **multilingual evaluation** of LLMs in **crisis relevant tasks** considering **stakeholders alignment**

Research Questions

RQ1

How does **Machine Translation performance** vary across linguistically diverse languages in the **crisis domain**?

RQ2

Do **LLMs align** with **human judgment** on urgency, and are they ***consistent*** across **languages**?

Experimental Setup

TICO-19 corpus

+20 typologically diverse languages

COVID-19 crisis domain
2k sentences / language

Human-translated gold standard

Experimental Setup

TICO-19 corpus

+20 typologically diverse languages

COVID-19 crisis domain
2k sentences / language

PubMed

Human-translated gold standard

Diverse MT *family* models

NLLB-200 — Meta

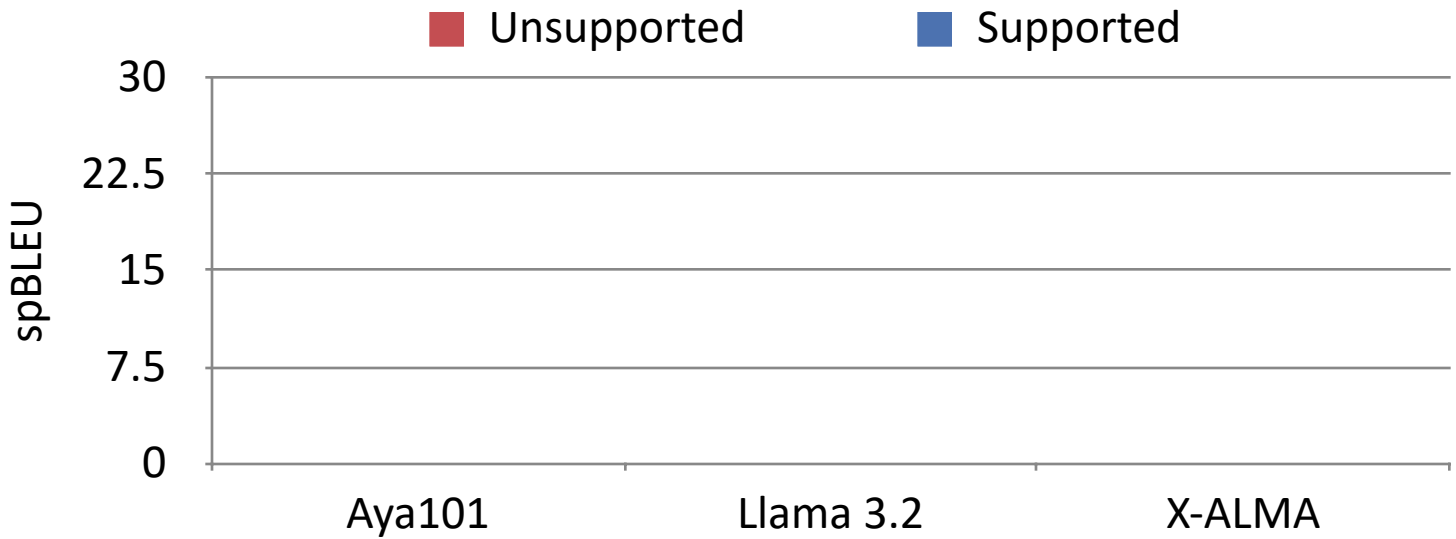
Aya-101 — Cohere For AI

X-ALMA

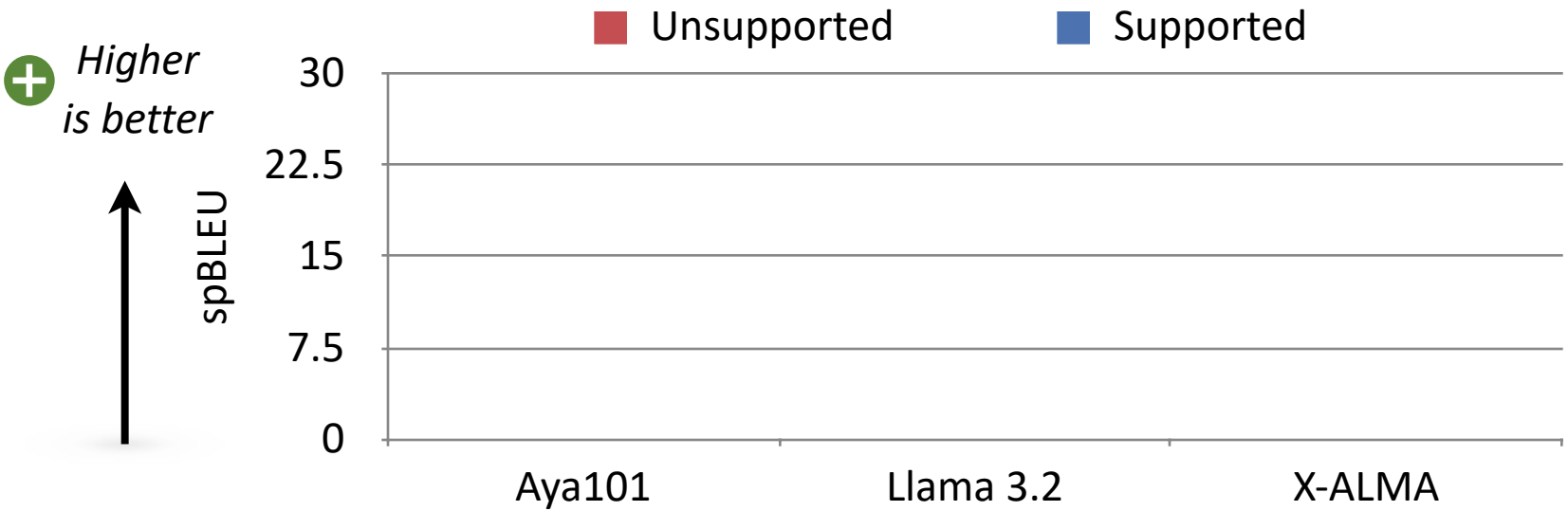
Llama 3.2 — Meta

MT quality ~ official language support

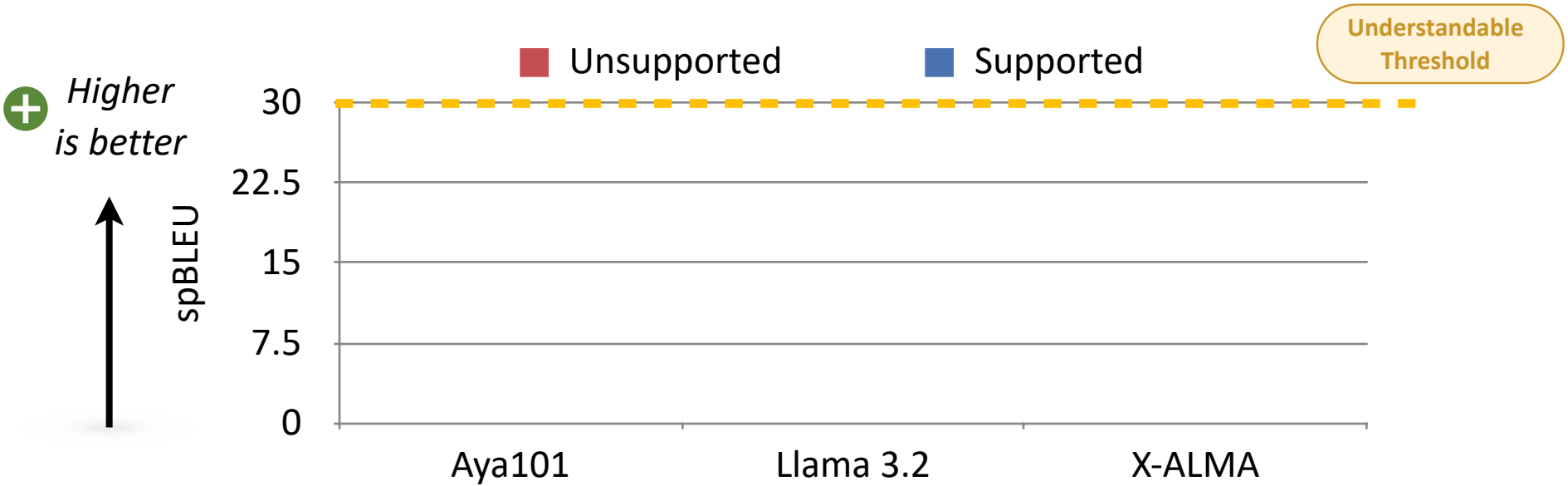
MT quality ~ official language support



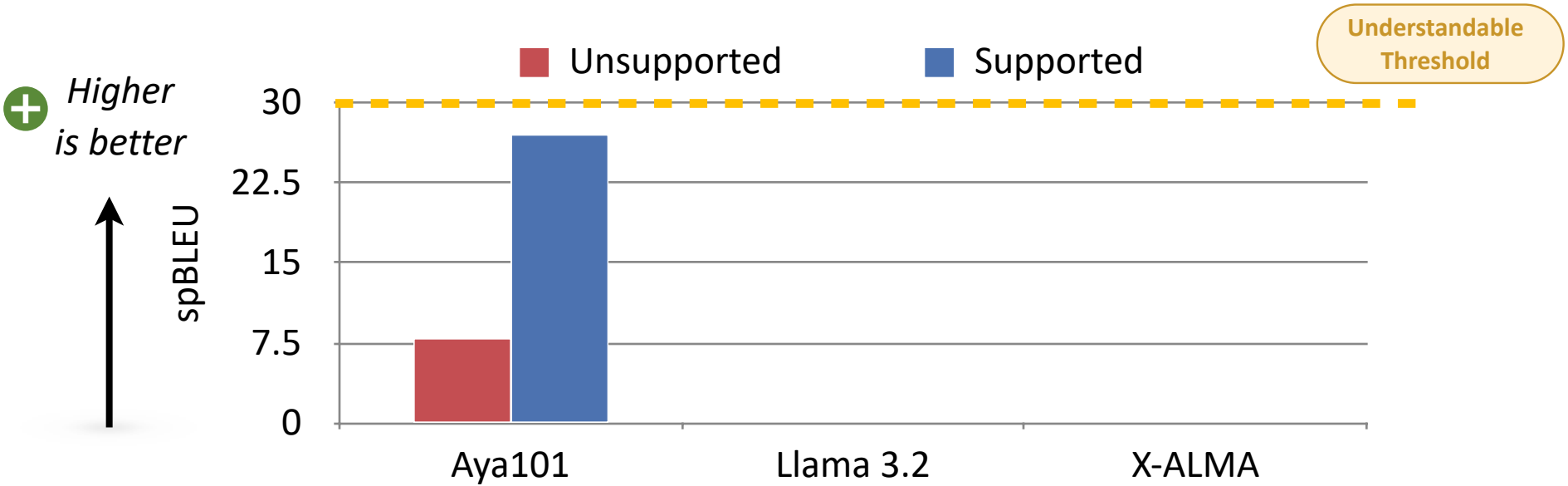
MT quality ~ official language support



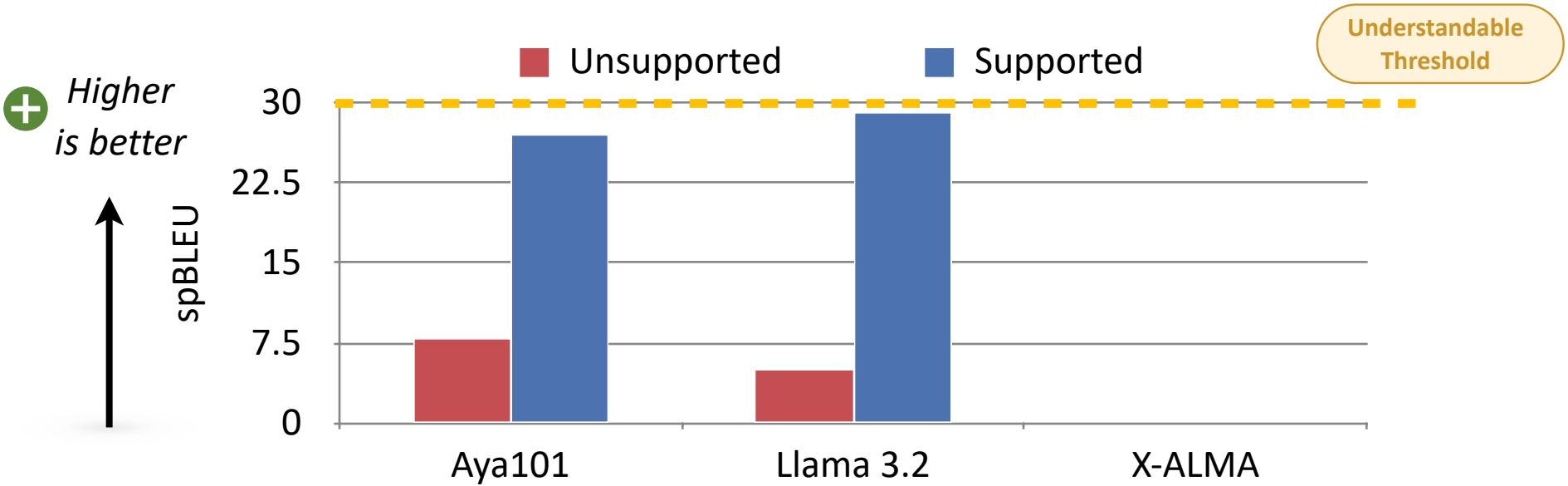
MT quality ~ official language support



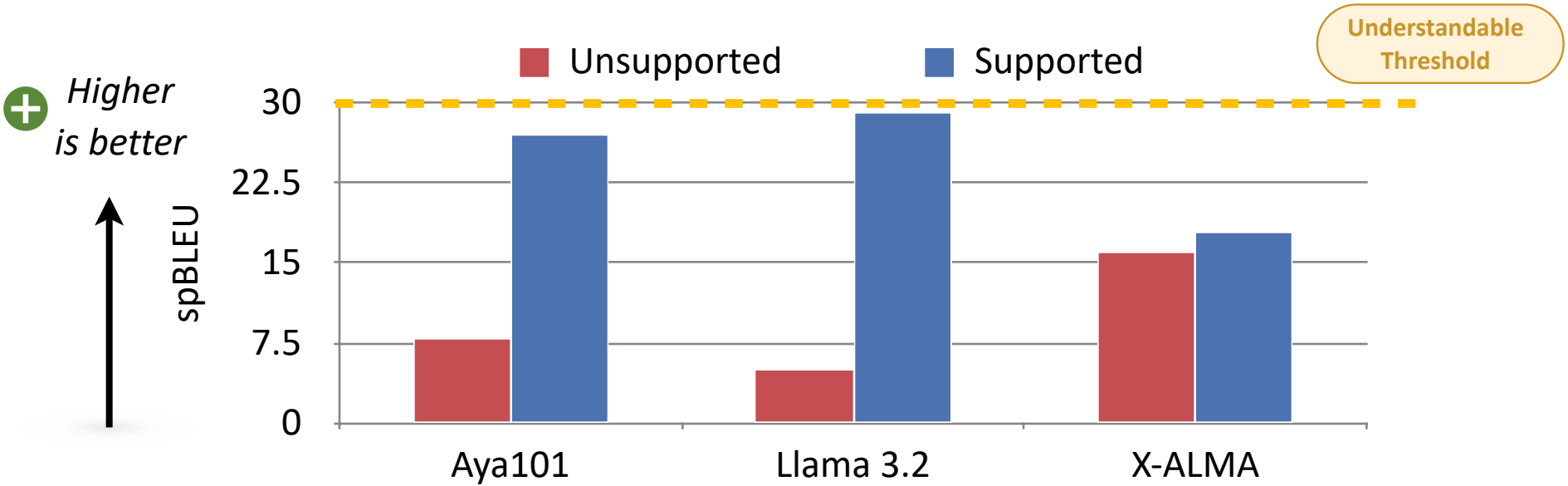
MT quality ~ official language support



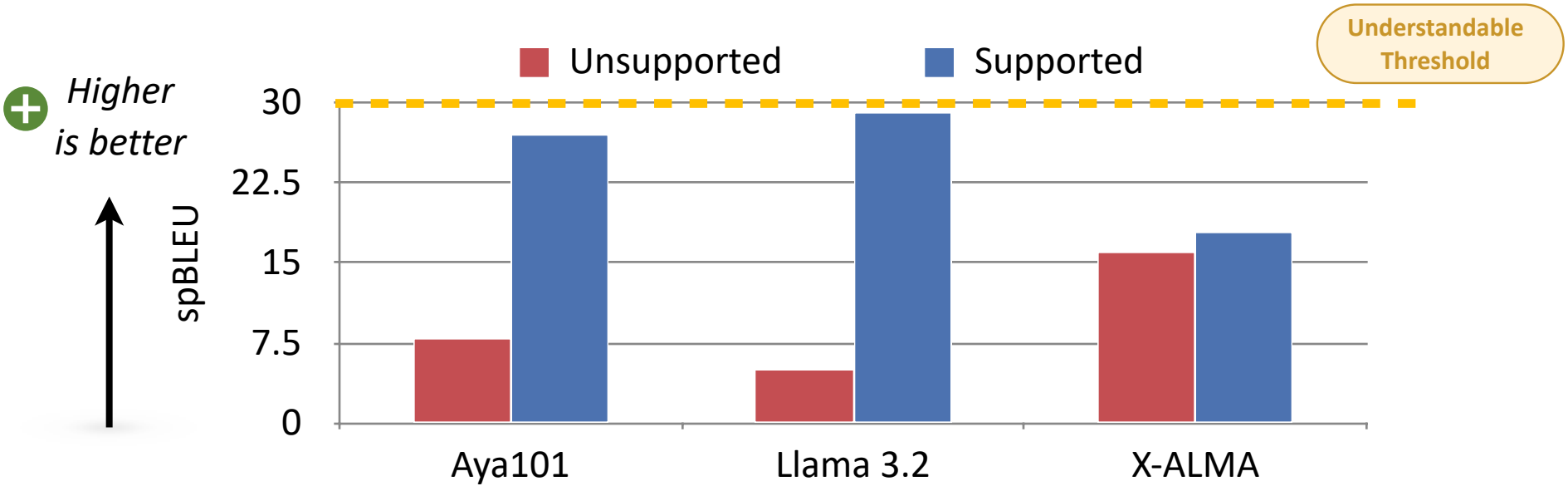
MT quality ~ official language support



MT quality ~ official language support



MT quality ~ official language support

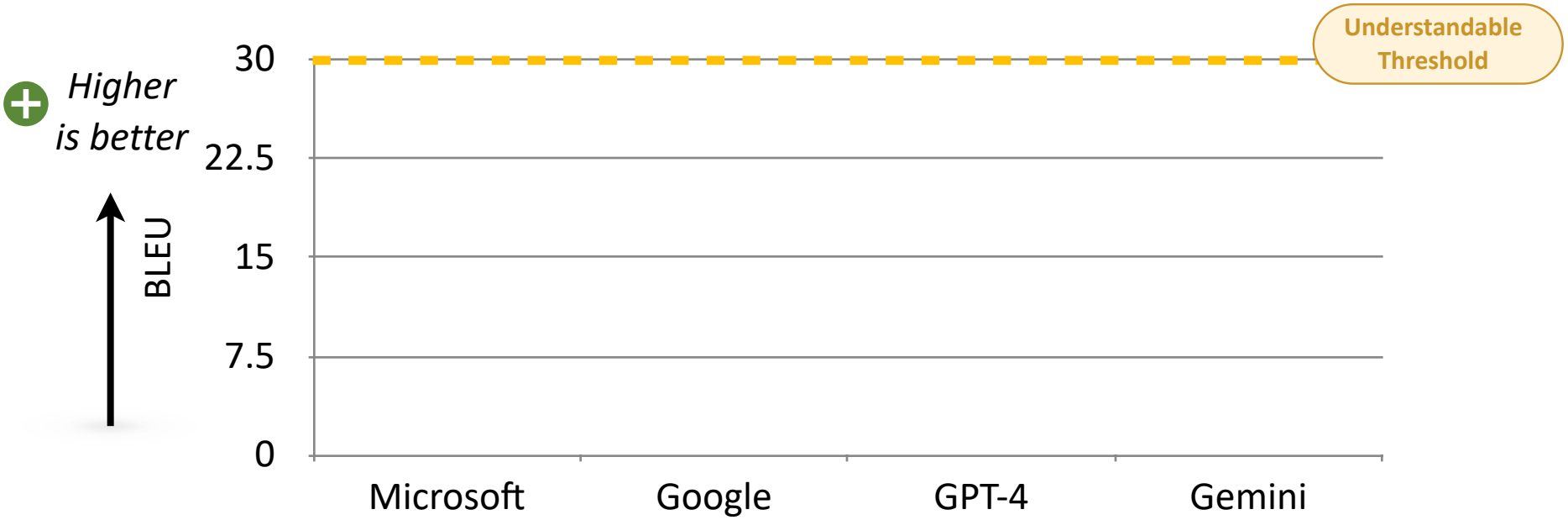


Takeaway Practical MT performance collapses for unsupported crisis languages.

Commercial Models Still Below the Bar for Usable Translation

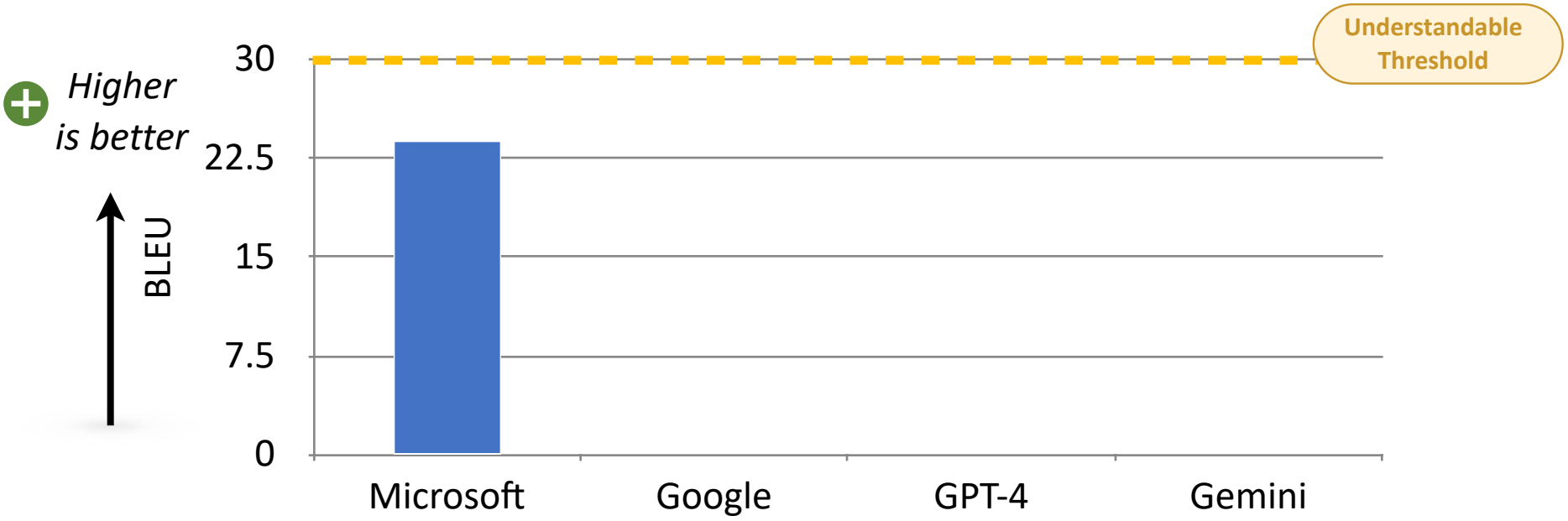
From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Commercial Models Still Below the Bar for Usable Translation



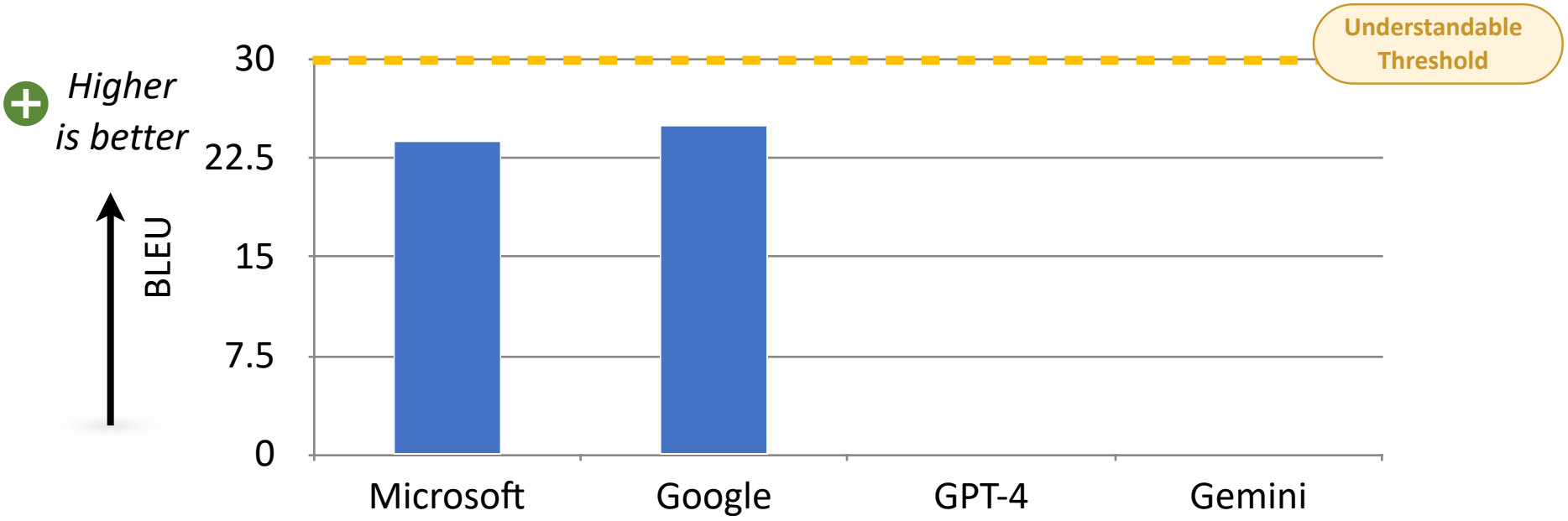
From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Commercial Models Still Below the Bar for Usable Translation



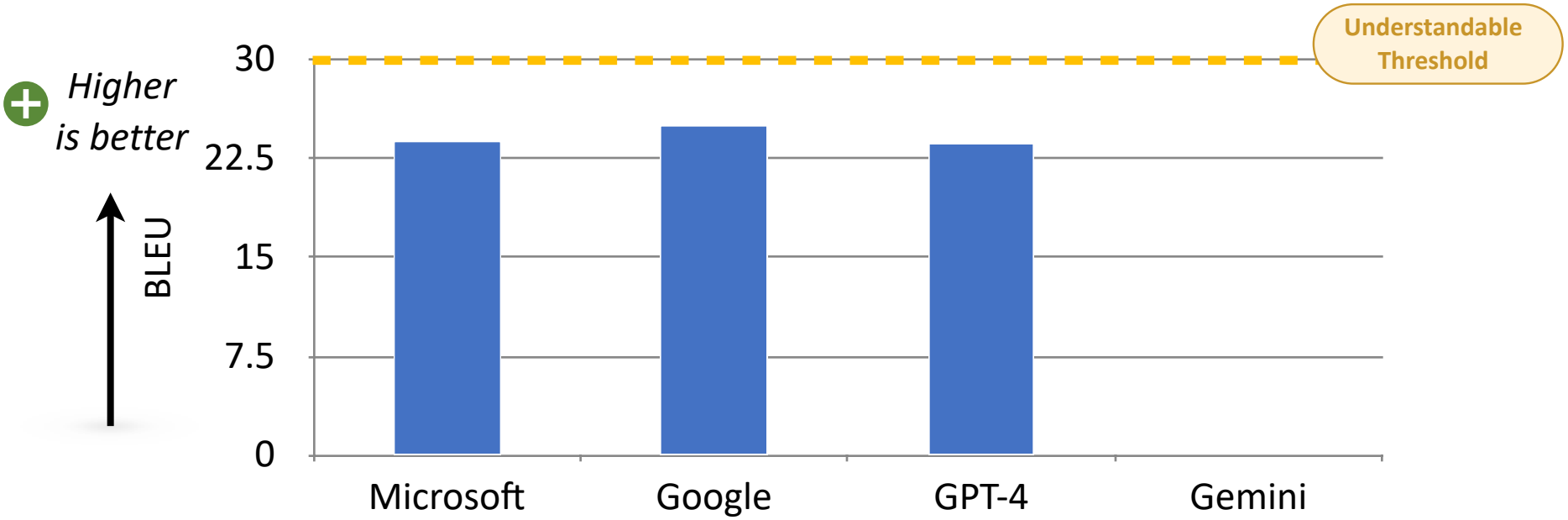
From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Commercial Models Still Below the Bar for Usable Translation



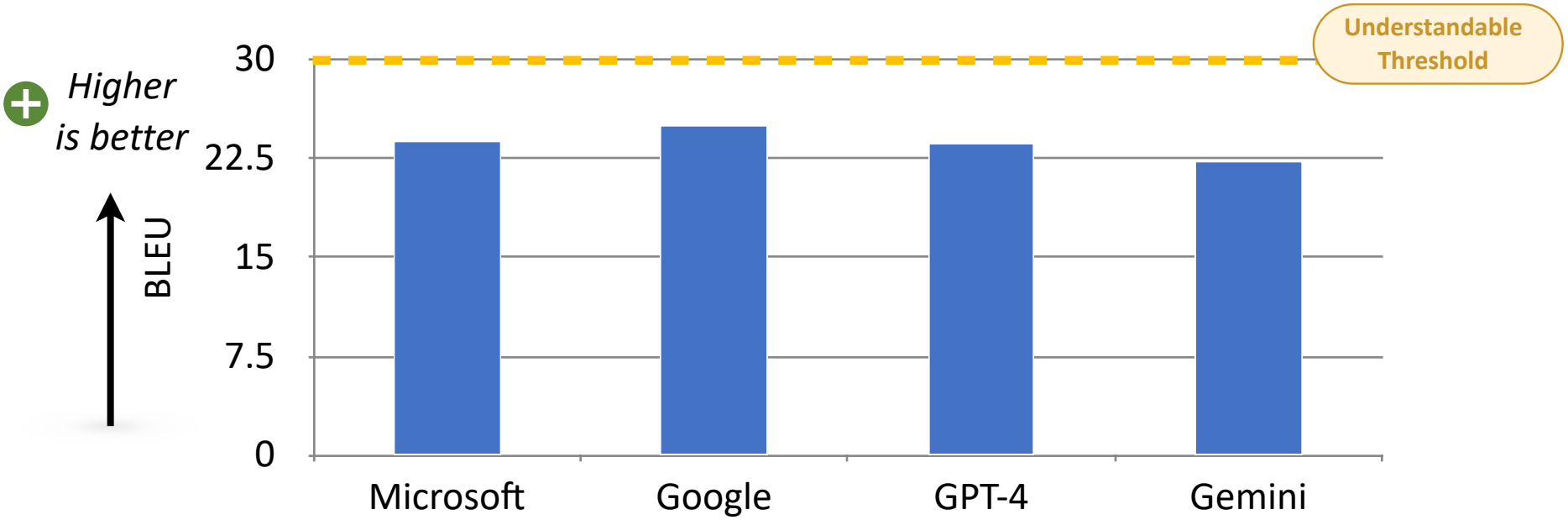
From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Commercial Models Still Below the Bar for Usable Translation



From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Commercial Models Still Below the Bar for Usable Translation



From Bansal et al. "Is MT Ready for the Next Crisis or Pandemic?" (2026)

Research Questions

RQ1

How does **MT performance** vary across linguistically diverse languages in the **crisis domain**?

RQ2

Do **LLMs align** with **human judgment** on urgency, and are they ***consistent*** across **languages**?

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Experimental Setup

1) GPT-4o

Generate messages convening different level of **urgency**

Natural disasters

Institutional Tone

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Experimental Setup

1) GPT-4o — Generate messages convening different level of **urgency**

- *Natural disasters*
- *Institutional Tone*

2) NLLB (Meta) **Translation**

19 languages

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Experimental Setup

1) GPT-4o — Generate messages convening different level of **urgency**

Natural disasters

Institutional Tone

2) NLLB Meta **Translation**

19 languages

Classify: **Level of urgency (1-5)**

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Experimental Setup

1) GPT-4o — Generate messages convening different level of **urgency**

Natural disasters

Institutional Tone

2) NLLB — Meta — **Translation**

19 languages

Classify: **Level of urgency (1-5)**

3a) ***LLM Annotations***

Llama 3.2 — Meta

Does Machine Translation Undermine LLM Reliability in Crisis Settings?

Same application, different languages, same results?

Lack of parallel data

Experimental Setup

1) GPT-4o — Generate messages convening different level of **urgency**

Natural disasters

Institutional Tone

2) NLLB Meta **Translation**

19 languages

Classify: **Level of urgency (1-5)**

3a) ***LLM Annotations***

Llama 3.2 Meta

3b) ***Human Annotations***

Same scenario, humans agree

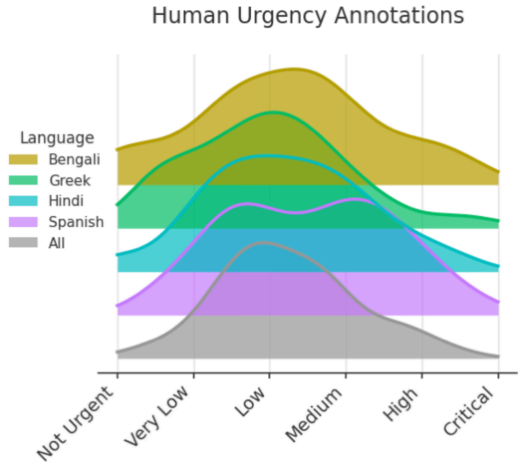
Source scenario (English) Human & LLM: HIGH

"All residents in the flood zone must evacuate immediately. Water levels are rising rapidly and the situation is becoming life-threatening."

Same scenario, humans agree

Source scenario (English) **Human & LLM: HIGH**

"All residents in the flood zone must evacuate immediately. Water levels are rising rapidly and the situation is becoming life-threatening."

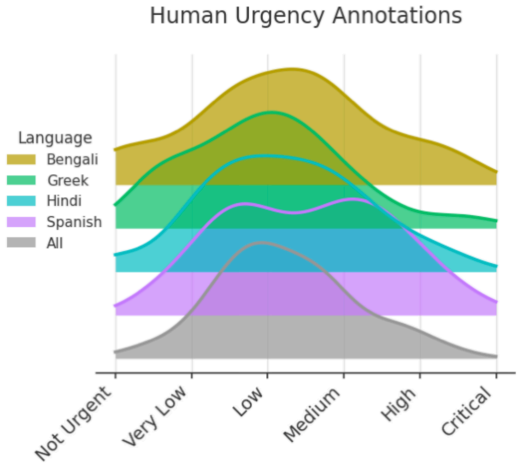


Same scenario, humans agree

Source scenario (English) **Human & LLM: HIGH**

"All residents in the flood zone must evacuate immediately. Water levels are rising rapidly and the situation is becoming life-threatening."

LANGUAGE	HUMANS
Span. (LA)	High ✓
Bengali	High ✓
Greek	High ✓
Hindi	High ✓

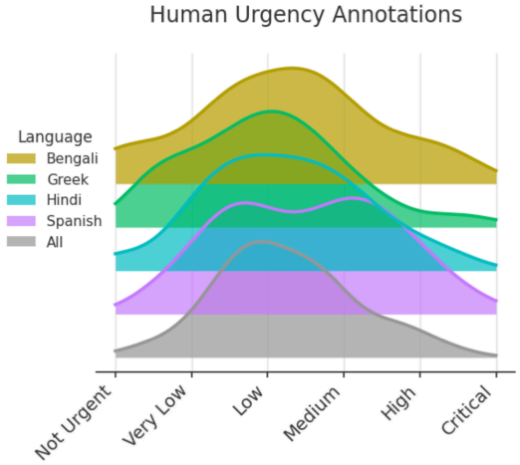


Same scenario, humans agree **but LLM don't**

Source scenario (English) Human & LLM: HIGH

"All residents in the flood zone must evacuate immediately. Water levels are rising rapidly and the situation is becoming life-threatening."

LANGUAGE	HUMANS	LLM LABEL
Span. (LA)	High ✓	Critical
Bengali	High ✓	Medium
Greek	High ✓	Very Low
Hindi	High ✓	Not Urgent



Llama 3.2 is inconsistent in urgency classification

en	*A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed inst…*	Critical
es-LA	*Se ha detectado una importante fuga de gas en la zona residencial; equipos de evacuación y reparación deben…*	High
id	*Kebocoran gas besar telah terdeteksi di daerah perumahan; tim evakuasi dan perbaikan harus segera dikerah…*	Critical
mr	*निवासी भागात मोठ्या प्रमाणात गॅस गळती झाल्याचे आढळून आले आहे; स्थलांतर आणि दुरुस्ती पथके तातडीने तैनात करणे आवश्यक…*	Low
zh	*在住宅区发现大气泄漏, 必须立即派出疏散和维修小组.*	Medium
am	*በመኖሪያ አካባቢ አንድ ትልቅ የጋዝ ፍሰሽ ተገኝቷል፤ የማስወገጃና የጥገና ቡድኖች ወዲያውኑ መሰማራት አ…*	Not Urgent

Llama 3.2 is inconsistent in urgency classification

en	*A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed inst…*	Critical
es-LA	*Se ha detectado una importante fuga de gas en la zona residencial; equipos de evacuación y reparación deben…*	High
id	*Kebocoran gas besar telah terdeteksi di daerah perumahan; tim evakuasi dan perbaikan harus segera dikerah…*	Critical
mr	*निवासी भागात मोठ्या प्रमाणात गॅस गळती झाल्याचे आढळून आले आहे; स्थलांतर आणि दुरुस्ती पथके तातडीने तैनात करणे आवश्यक…*	Low
zh	*在住宅区发现大气泄漏, 必须立即派出疏散和维修小组.*	Medium
am	*በመኖሪያ አካባቢ አንድ ትልቅ የጋዝ ፍሰሽ ተገኝቷል፤ የማስወገጃና የጥገና ቡድኖች ወዲያውኑ መሰማራት አ…*	Not Urgent

Llama 3.2 is inconsistent in urgency classification

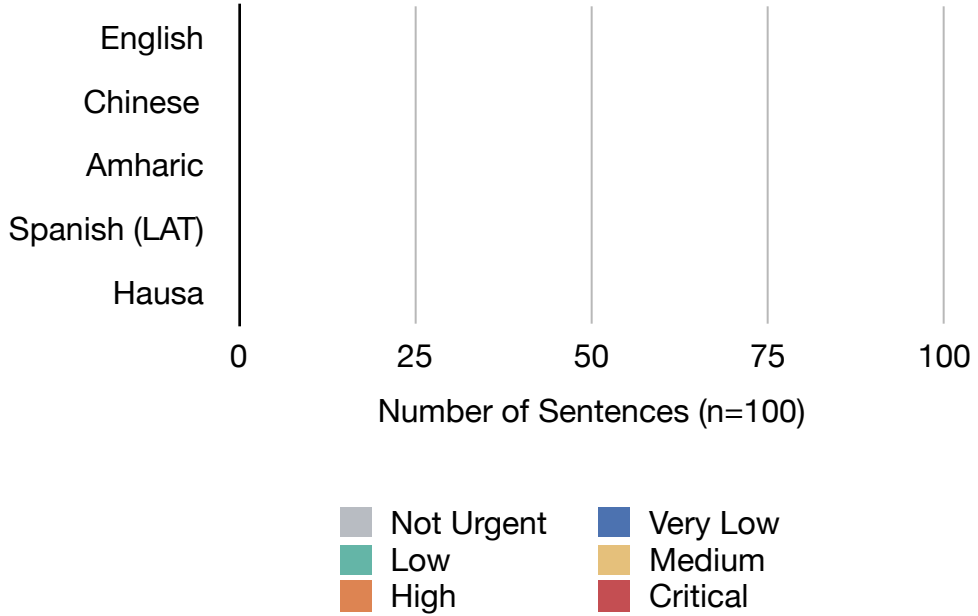
en	*A major gas leak has been detected in the residential area; evacuation and repair teams must be deployed inst…*	Critical
es-LA	*Se ha detectado una importante fuga de gas en la zona residencial; equipos de evacuación y reparación deben…*	High
id	*Kebocoran gas besar telah terdeteksi di daerah perumahan; tim evakuasi dan perbaikan harus segera dikerah…*	Critical
mr	*निवासी भागात मोठ्या प्रमाणात गॅस गळती झाल्याचे आढळून आले आहे; स्थलांतर आणि दुरुस्ती पथके तातडीने तैनात करणे आवश्यक…*	Low
zh	*在住宅区发现大气泄漏, 必须立即派出疏散和维修小组.*	Medium
am	*በመኖሪያ አካባቢ አንድ ትልቅ የጋዝ ፍሳሽ ተገኝቷል፤ የማስወገጃና የጥገና ቡድኖች ወዲያውኑ መሰማራት አ…*	Not Urgent

LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***

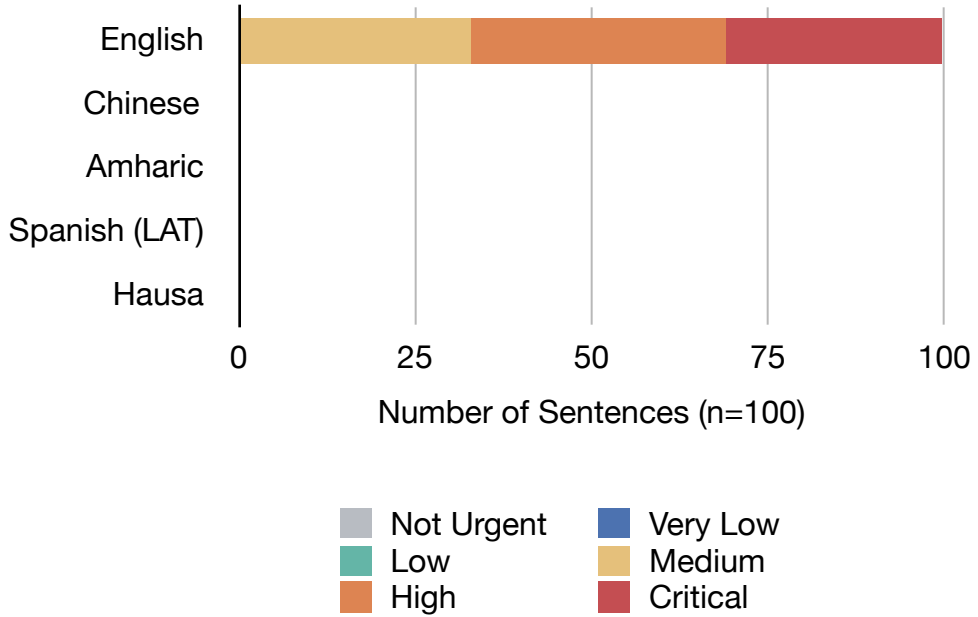
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



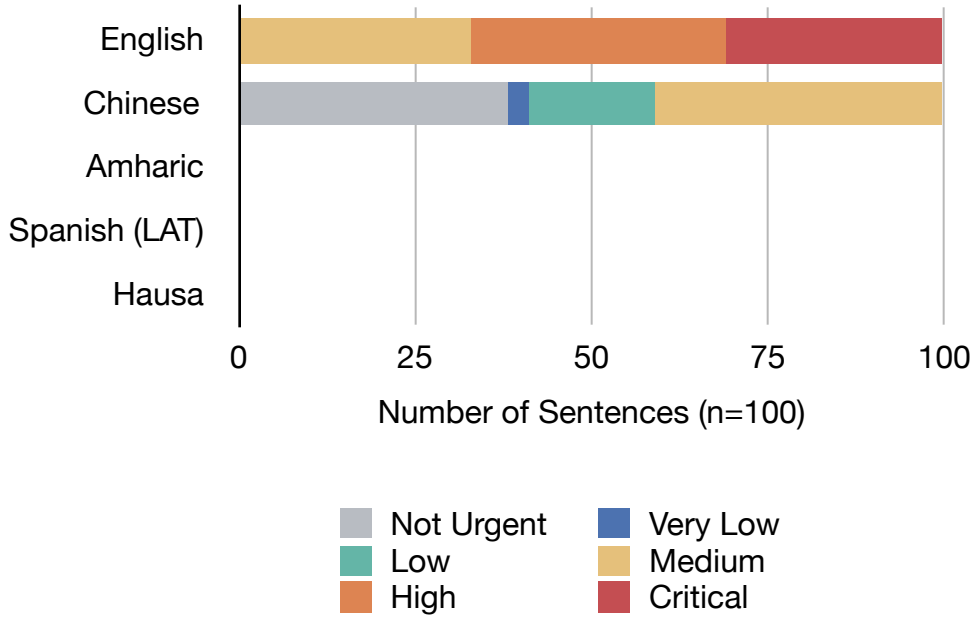
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



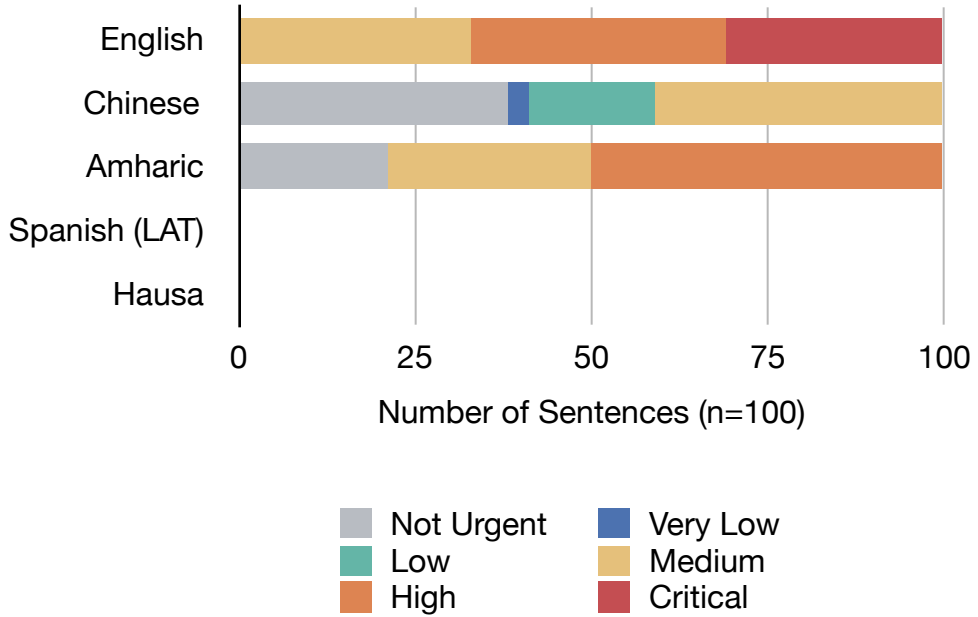
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



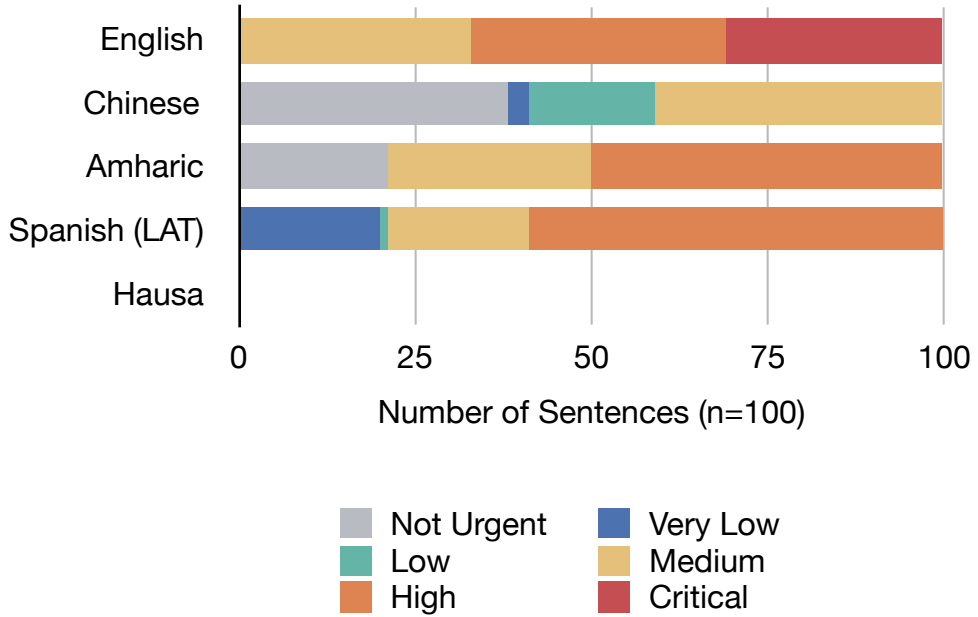
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



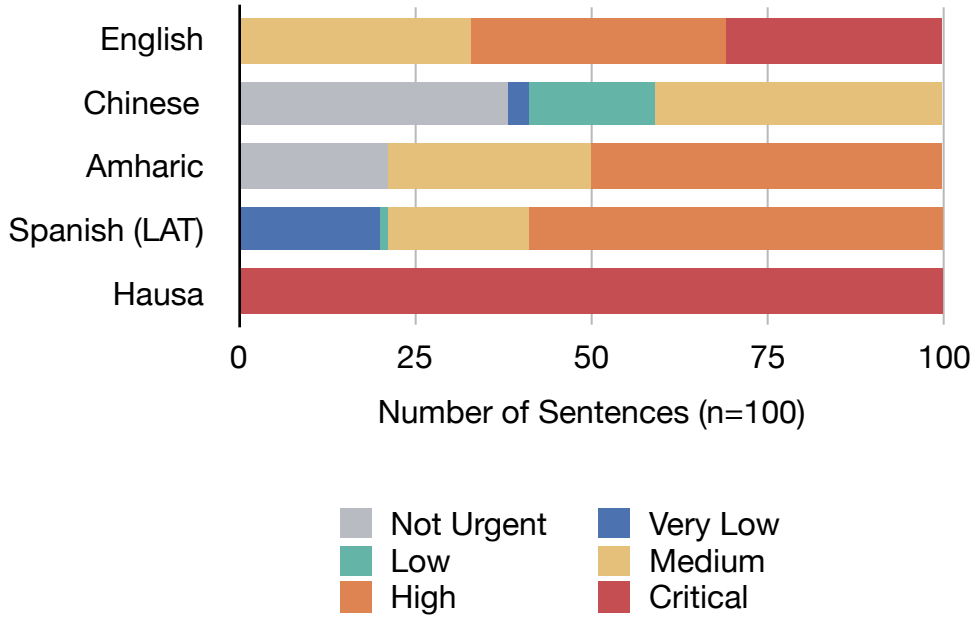
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



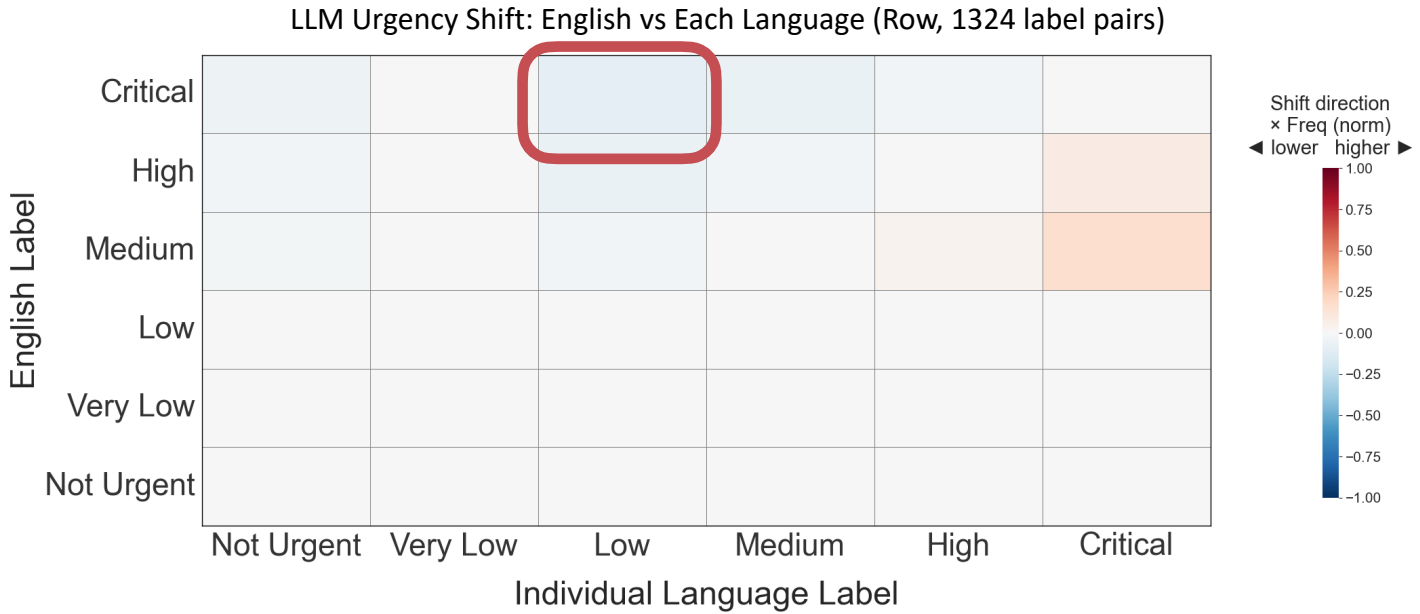
LLMs downgrade crisis urgency in translation

Urgency *label distribution* ***differs*** *significantly* ***across languages***



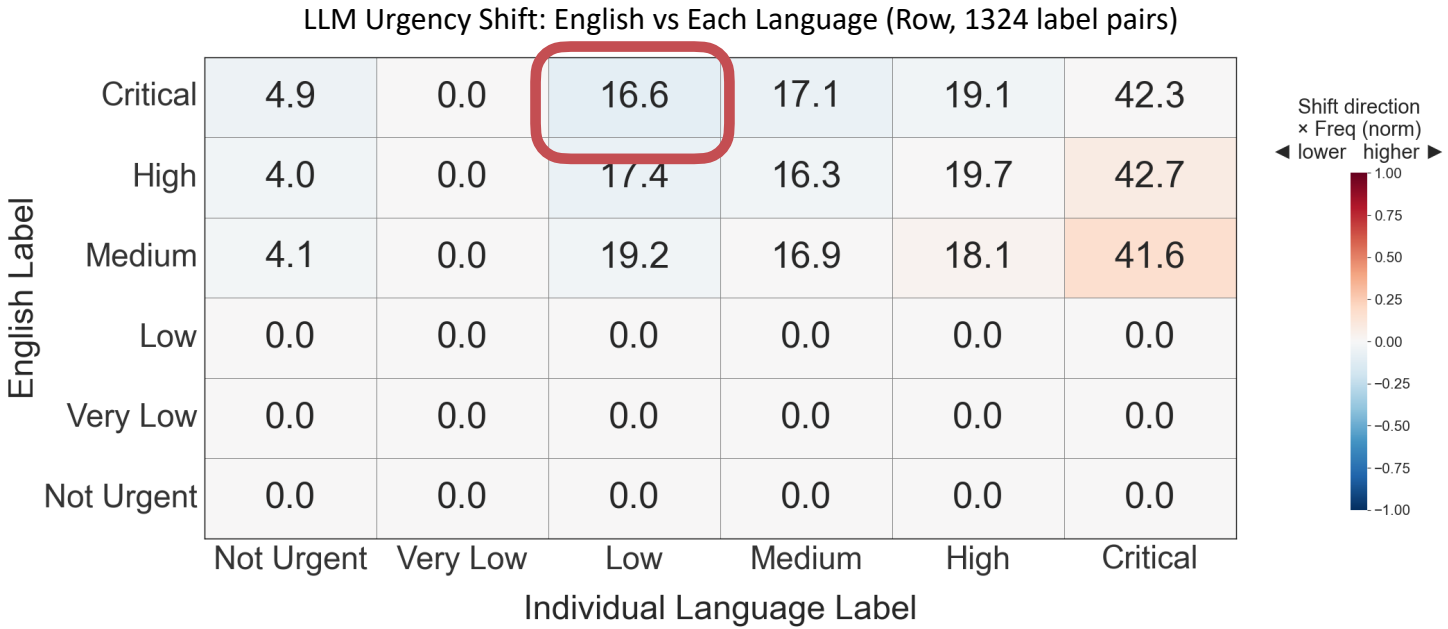
How Urgency shifts are downward/upward?

% of sentences labeled as **Critical** in English that are labeled as **Low** in X language



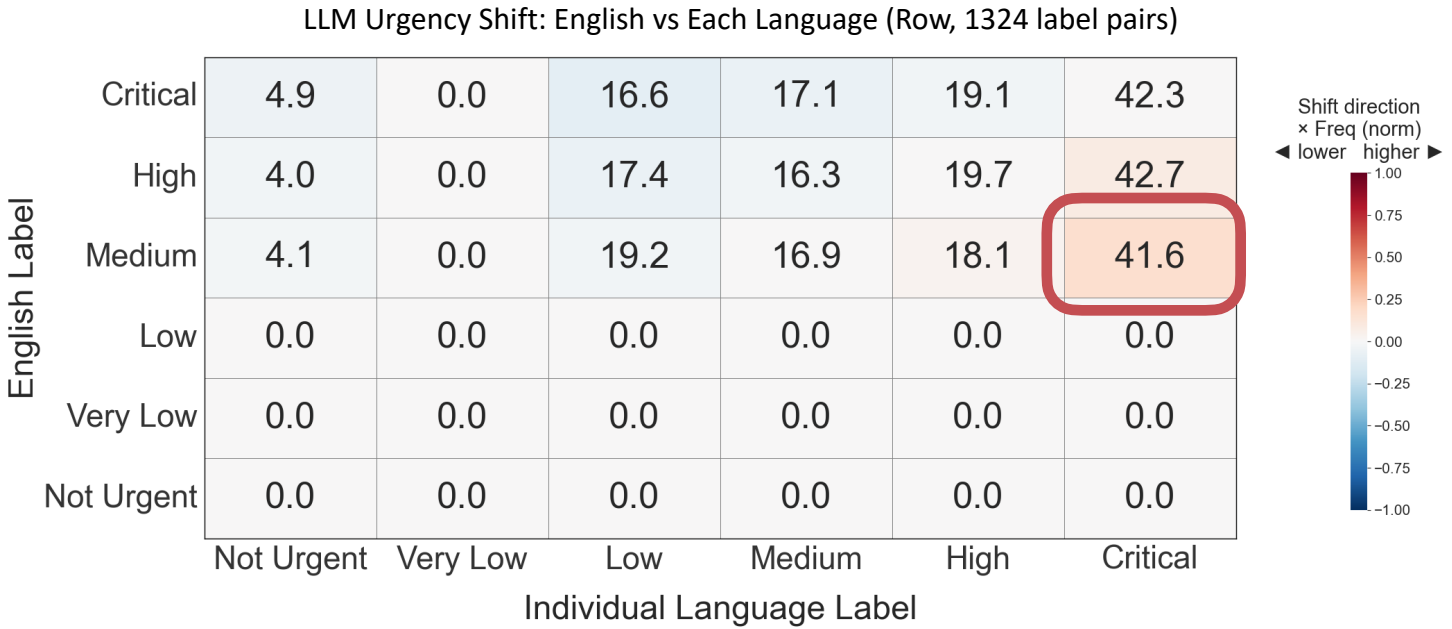
How far does urgency levels drop after translation?

% of sentences labeled as **Critical** in English that are labeled as **Low** in X language



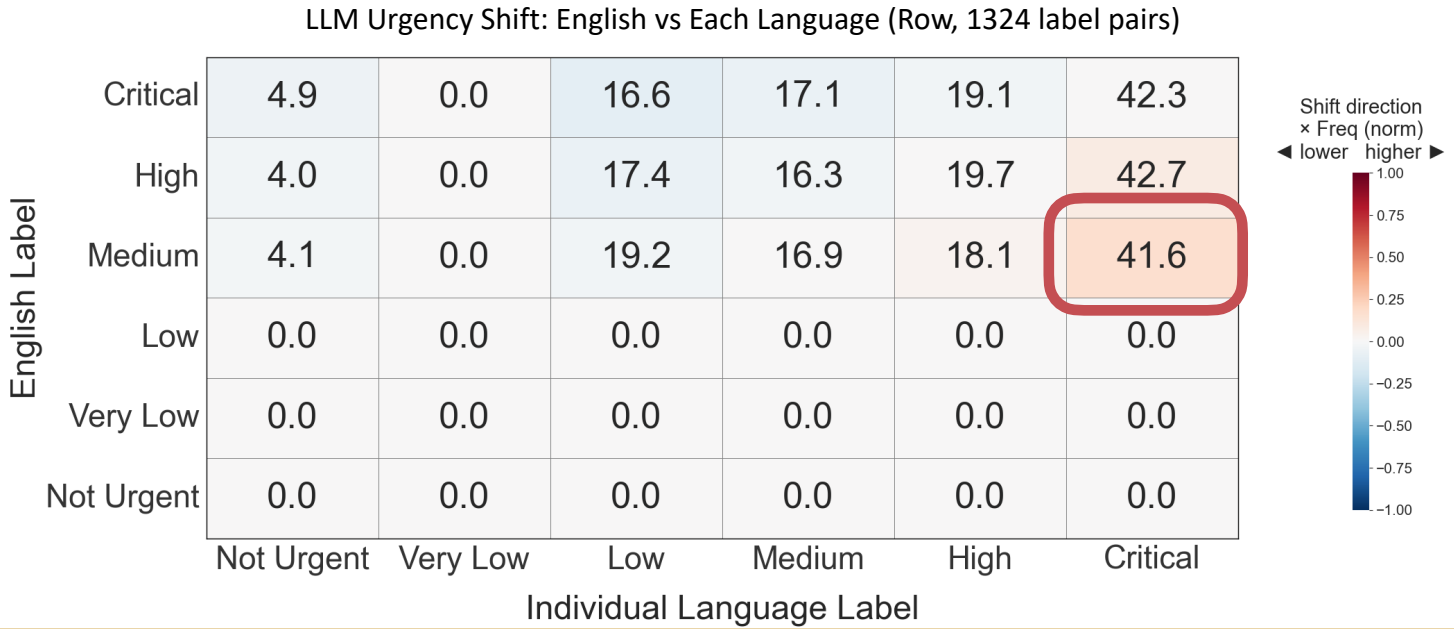
How far does urgency levels drop after translation?

% of sentences labeled as **Critical** in English that are labeled as **Low** in X language



How far does urgency levels drop after translation?

% of sentences labeled as **Critical** in English that are labeled as **Low** in X language



Takeaway **LLMs perform urgency classification inconsistently across languages**

Preliminary Findings

RQ1

How does **MT performance** vary across linguistically diverse languages in the **crisis domain**?

*Yes, overall **practical MT** performance **collapses** for **unsupported** languages*

RQ2

Do **LLMs align** with **human judgment** on urgency, and are they ***consistent*** across **languages**?

*No, **LLM do not align with humans** and even more concerningly they perform **inconsistently***

Future Work

Future Work

- *Use of real life world data (looking for collaboration!)*

Future Work

- *Use of real life world data (looking for collaboration!)*
- *Explore other relevant classification tasks (e.g. severity assessment)*

Future Work

- *Use of real life world data (looking for collaboration!)*
- *Explore other relevant classification tasks (e.g. severity assessment)*
- *More models for translation and annotation (open weights vs commercial)*

Future Work

- *Use of real life world data (looking for collaboration!)*
- *Explore other relevant classification tasks (e.g. severity assessment)*
- *More models for translation and annotation (open weights vs commercial)*
- *Alignment between LLMs and different stakeholders*

Thank you

Work supported by a 3 year NSF grant.
Find more at **LT4CPR.github.io**

I'm happy to collaborate in building and evaluating AI models for crisis settings

Call for papers!

Belu Ticona
mticonao@gmu.edu

Antonios Anastasopoulos
antonis@gmu.edu

Speech+NLP+HCI

SCAN TO LEARN MORE

GEORGE MASON UNIVERSITY