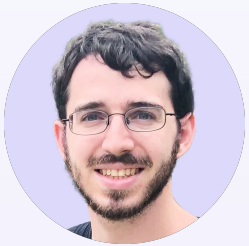
The path to LLM adoption: Key use cases, benchmarks, and trust-building strategies

Alex Yanishevsky
Senior Director, AI Solutions, Smartling



Jason Rauchwerk
MT Deployment Engineer, Smartling

global ready conference by SMARTLING

Agenda

- How are LLMs performing today?
- Deep dive: New and updated use cases
- Opportunities for DEI
- Building trust

How are LLMs performing today?

Improvements in AI Post-Processing

- Updates on fuzzy match repair
- Glossary term insertion
- Smoothing
- EEE (now LQE)

- How to best utilize AI Post-Processing

AI Toolkit Recent Improvements:
Fuzzy Match Repair

- **Fuzzy match repair and new smoothing prompt:** Introducing more robust RAG, examples from TMs in the prompt, including glossary entries explicitly. This should improve glossary issues for both FMR and machine translation smoothing.
- **State-of-the-art LLM:** Gemini Flash 2.0 is live for FMR and is more accurate than the previous generation.
- **Added threshold:** Everything that is 95% and higher is not subject to FMR.
- **Fuzzy match repair unnatural phrasing or worse than TM match:** This could be due to hallucinations or unreliable data in the TM. This can happen more on strings with numeric values and date/time. In general, this should happen in very few strings relative to all strings being processed.

AI Toolkit Dashboard: FMR Improvements

- 25.1% improvement over original match
  - 60% original match -> 85% fuzzy match
  - Dec 2024-Mar 2025

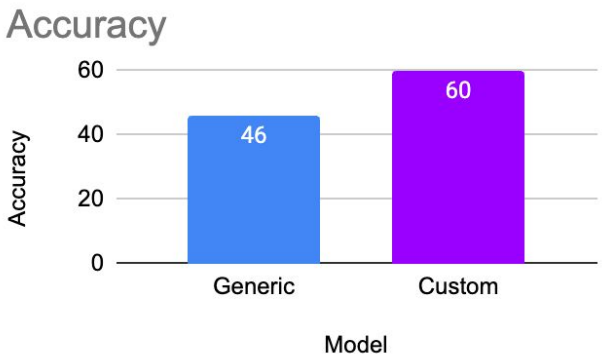
Source

AI Toolkit recent improvements:
Formality, GTI and Smoothing

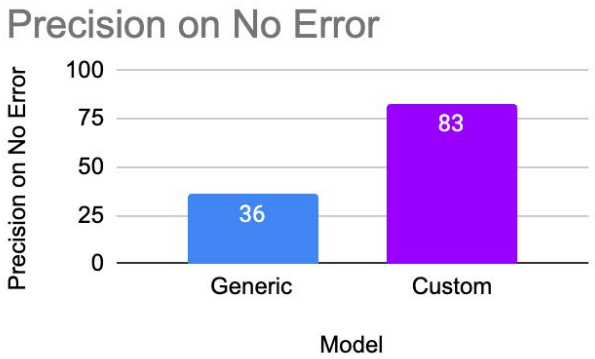
- **Formality Handling:** We have migrated to Gemini version, 1.5v2 Flash. In our research, using the same prompt is yielded improvements. The prompt contains instructions such as “do not go into pejorative or casual.”
- **Glossary Term Insertion:** GTI works with 80-90% accuracy depending on languages at the moment. Moving to Gemini Flash 2 will improve accuracy slightly. We have modified our prompt and research has shown additional improvement in accuracy.
- **Smoothing prompt** has been dramatically improved. We are testing LLMs other than GPT 4o to see if we get better benchmarks.

AI Toolkit recent improvements: LQE

Customized (Fine tuned) LQE
models are more accurate than generic LQE models by 30% or more



Precision
Of all the instances predicted as positive, how many are actually positive?



Best AI results for a customer

- No linguistic assets

- Somewhat consistent linguistic assets
- Meager leverage from TM
- Autoselect

- Clean, consistent, and reliable linguistic assets
- Lots of leverage from TM
- Trained MT engines

Deep dive: New and updated use cases

NMT vs LLM: MT use case

NMT

- More literal
- Better adherence to brand terminology
- For human-in-the-loop, less creative content

LLM

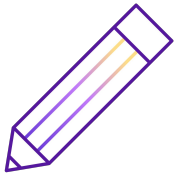
- More fluent and natural-sounding
- Lesser adherence to brand terminology
- For no human-in-the-loop, more creative content

Switch?

- **Use both based on content?**
- NMT for HITL, less creative content
- LLM for no HITL, more creative content

LLMs: Other use cases

Where can AI fit in the Translation space?

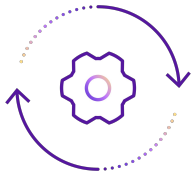


Source editing

Editing source content for higher quality translations (e.g. UGC)

LLM Translations

Translating directly with AI technology, like LLMs, bypassing the use of MT



Translation augmentation

Augmenting MT translations through the use of AI during and post translation

Localized content creation

Death to source, translations are generated directly with AI

Automated post-editing increases quality by 4-5 MQM

Benchmarking LLMs: Evaluation versus Estimation

Evaluation:
You have a reference

Source: Find Express Service Code
Reference: Encuentre el código de servicio expreso
MT: Buscar código de servicio exprés

Goal: How closely does MT match the reference? How little do I have to edit?

Estimation:
You do NOT have a reference.

Source: Find Express Service Code
MT: Buscar código de servicio exprés

Goal: Is the MT a good translation of the source?

Autoscoring

Source string

These documents can sometimes be needed to satisfy the requirements of the customs authorities or other government bodies and agencies that have a vested interest in what leaves and enters their country.

Human

A veces, estos documentos pueden ser necesarios para satisfacer los requisitos de las autoridades aduaneras u otros organismos y agencias gubernamentales que tienen un interés personal en lo que sale y entra en su país.

Trained machine translation

Estos documentos **a** veces pueden ser necesarios para satisfacer los requisitos de las autoridades aduaneras u otros organismos y agencias gubernamentales que tienen un interés creado en lo que sale y entra en su país.

BLEU = 79.96 | **Edit distance =** 6.66

Semantic scoring: Beyond textual scoring

When textual scoring fails

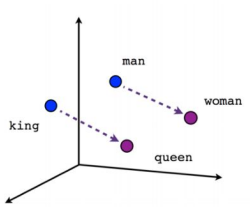
Let's eat, Grandmother!
Let's eat Grandmother

Ref	Hyp	Words	BLEU	PE Distance	TER %	CharacTER	ChrfScore
After lengthy deliberations, the jury found the defendant guilty	After lengthy deliberations, the jury found the defendant **not** guilty	11	89.27	0.06	0.1	0.06	95.75
Insert the oil extractor tube and pour gas into the engine	Insert the oil extractor tube and pour **oil** into the engine	12	93.75	0.05	0.09	0.05	87.75

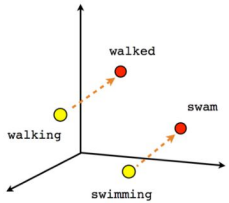
Semantic scoring: Beyond textual scoring 2

Word embeddings

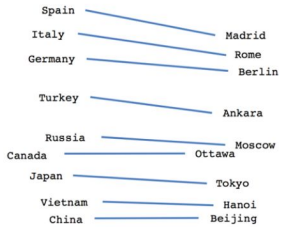
1. Numerical representation of words
2. Words that are similar are plotted close to each other



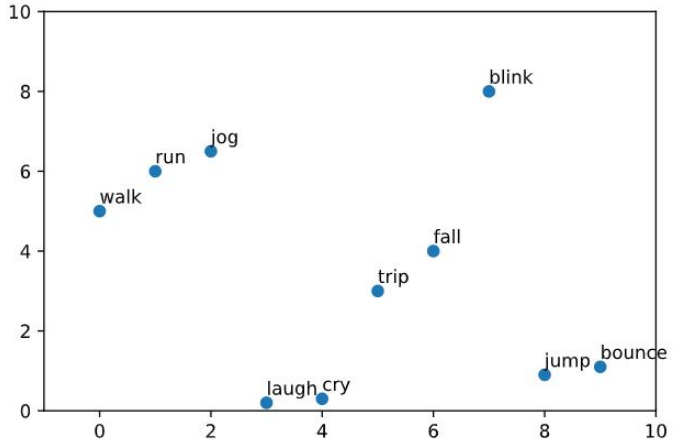
Male-Female



Verb tense



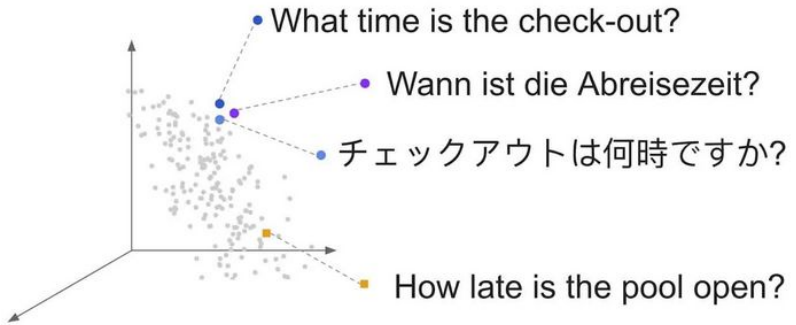
Country-Capital



Semantic scoring: Beyond textual scoring 3

Sentence Embeddings

1. Expand concept of word embeddings to sentences
2. Consider source and target

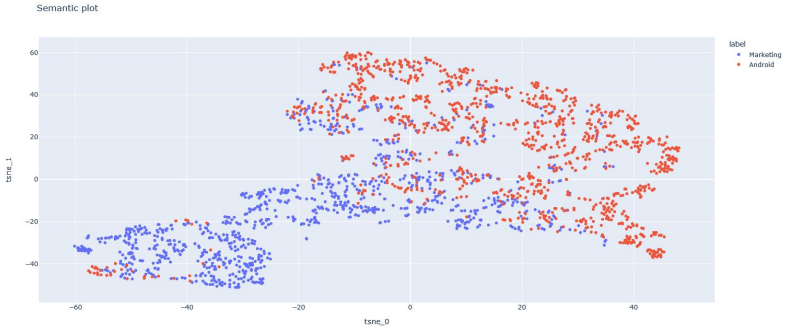


Semantic scoring: Beyond textual scoring 4

Sentence Embeddings

1. Lower scores indicates source and target may not match semantically
2. Erroneous information for user
3. Information may not adhere to be brand tone or style

EN	FR	distiluse-base-multilingual-cased-v1_cosine_sim	laser_cosine_sim	LaBSE_cosine_sim
Hi Alyssa, We understand your concern that you did not receive an item from this restaurant.	Bonjour Florence, Nous comprenons votre inquiétude de ne pas avoir reçu un article de ce restaurant.	0.691071272	0.862825155	0.809140086
Hi Alyssa, We understand your concern that you did not receive an item from this restaurant.	Quelle est l'heure de départ	-0.094488777	0.279111385	0.174875483
I apologize for the inconvenience it has caused.	Je m'excuse pour le désagrément que cela a causé.	0.915562093	0.938958287	0.918583035



Benchmarking trained NMT vs fine-tuned LLM

In most cases, Trained NMT is still the better engine using both automatic textual scoring and semantic scoring metrics

Engine	Language	BLEU	Edit distance	Meteor	Comet	MetricX
AutoML	FR	63.07	9.18	82.96	0.92	1.41
PETLLM Full	FR	53.73	13.46	77.40	0.90	1.52
PETLLM_Full_Model2	FR	53.04	13.85	77.33	0.90	1.47
PETLLM_5K_Model2	FR	50.00	15.21	75.50	0.89	1.50
AutoML	PTBR	61.67	8.10	84.28	0.93	1.36
PETLLM Full	PTBR	50.16	12.68	77.46	0.92	1.52
PETLLM5k_model2	PTBR	44.74	16.13	73.07	0.91	1.53
PETLLM_Full_Model2	PTBR	49.12	13.73	75.43	0.92	1.48

Opportunities for DEI

AI gender debiasing and inclusive language

- **What?**
  Ensure that your translated content is debiased and inclusive, such as with occupational nouns.
- **Benefit:**
  Ensure that your translated content correctly references the addressed audience.

AI gender debiasing and inclusive language

Source Text

The governor is not in the office.

Traditional

Without LLM

El gobernador no está en la oficina.

LLM

With LLM

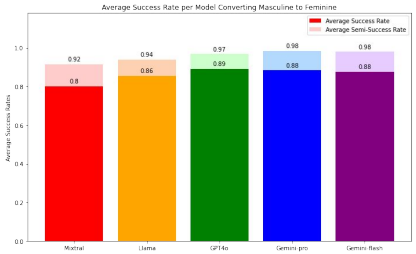
La gobernadora no está en la oficina.

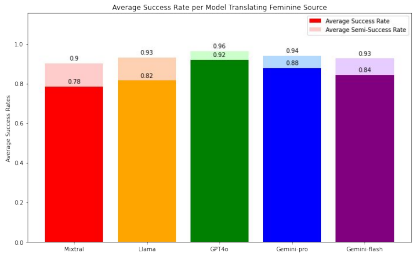
Pronoun conversion experiment

Method 1: Masculine Source → Masculine Translation → Nonbinary/Feminine Translation

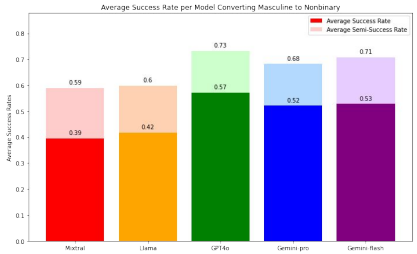
Method 2: Masculine Source → Nonbinary/Feminine Source → Nonbinary/Feminine Translation

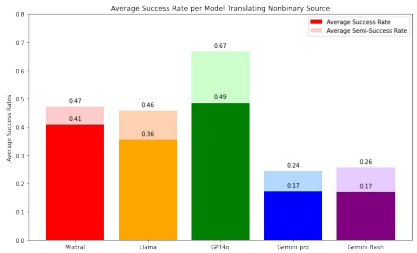
Feminine





Non-binary





AI gender debiasing and inclusive language prompt

Male to nonbinary (SNIPPET)

```
Create a gender-neutral version of the following Spanish text using these techniques:
1. Use generic nouns (e.g., change "Es un gran logro para el hombre" to "Es un gran logro
para la humanidad")
2. Use neutral pronouns (e.g., replace "Al interesado" with "A quien le pueda interesar")
3. Substitute "ser" with "estar" or rephrase (e.g., change "Estamos muy contentos con
este resultado" to "Este resultado nos alegra mucho")

Examples:
Input: ```Los médicos del Hospital La Paz han recomendado un nuevo tratamiento para el
dolor crónico que que promete mejorar la vida de pacientes con enfermedades avanzadas.```
Output: ```El personal médico del Hospital La Paz ha recomendado un nuevo tratamiento
para el dolor crónico que promete mejorar la vida de pacientes con enfermedades
avanzadas.```
```

Pronoun conversion experiment conclusions

Feminine conversion
- Over 80% success for all models
- Ready for production

Best performing nonbinary conversion
- Total success rate of 57%
- Not ready for use

Future progressions
- Evolution of LLMs
- Progression of nonbinary inclusivity in language

Future experiments/developments
- Changing terminology to be more appropriate

Building trust in LLMs

LLMs: A precocious youngster

Who is completely unpredictable...

English	German
But enterprises run internal networks that connect thousands of devices to each other and to the Internet.	Aber Unternehmen betreiben interne Netzwerke, die Tausende von Geräten miteinander und mit dem Internet verbinden.
I want to be able to give names to generic works, AND to particular translations AND to particular versions.	**Um generische Werke, bestimmte Übersetzungen und bestimmte Versionen zu benennen, können Sie folgendermaßen vorgehen:** **1. **Generische Werke benennen:**** **- Verwenden Sie den Originaltitel des Werks.** **- Fügen Sie den Namen des Autors hinzu, um Verwechslungen zu vermeiden.** **...**
Our solution protects against bot attacks while allowing verified bots through.	Unsere Lösung schützt vor Bot-Angriffen und lässt gleichzeitig verifizierte Bots durch.

Huh???

English	German
But enterprises run internal networks that connect thousands of devices to each other and to the Internet.	Aber Unternehmen betreiben interne Netzwerke, die Tausende von Geräten miteinander und mit dem Internet verbinden.
I want to be able to give names to generic works, AND to particular translations AND to particular versions.	**To name generic works, specific translations, and specific versions, you can follow these steps:** **1. **Name generic works:**** **- Use the original title of the work.** **- Add the author's name to avoid confusion.** **...**
Our solution protects against bot attacks while allowing verified bots through.	Unsere Lösung schützt vor Bot-Angriffen und lässt gleichzeitig verifizierte Bots durch.

How much autonomy?

Identifying hallucinations: Naive approaches

1.

Source to target length ratio

2.

Semantic similarity

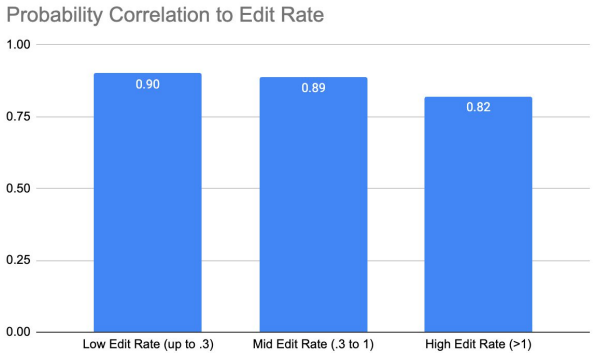
3.

LLM prompt:
If you're not certain, reply with “I don't know” as part of the reply

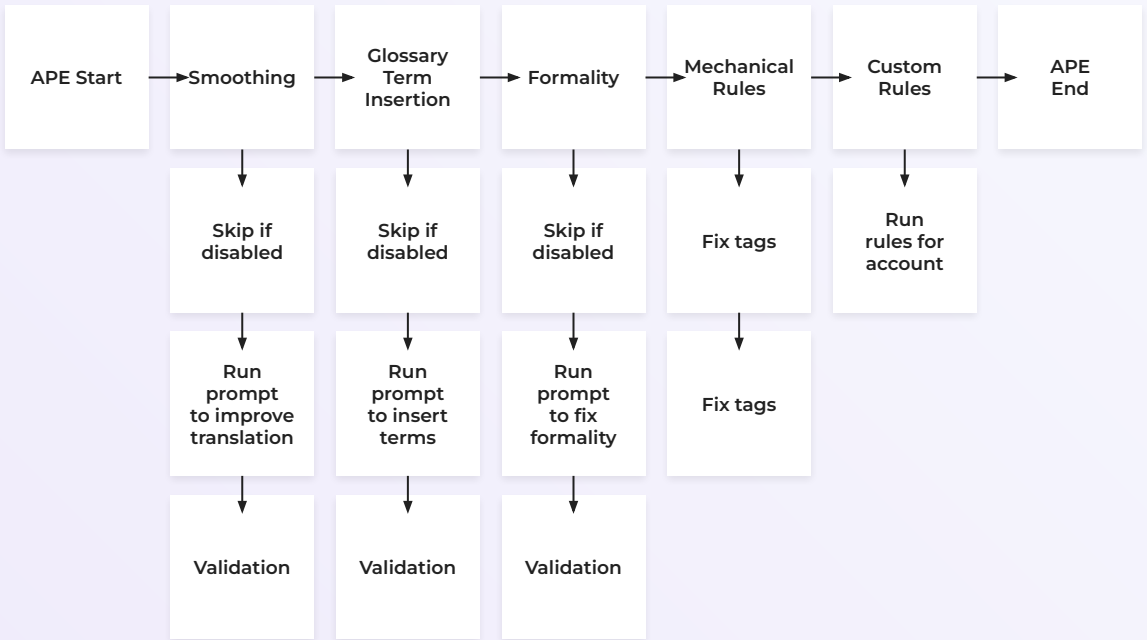
Log probability for each word

- Do we see peaks and valleys?
- How profound is the variance?
- Can we correlate to expected edit rate?

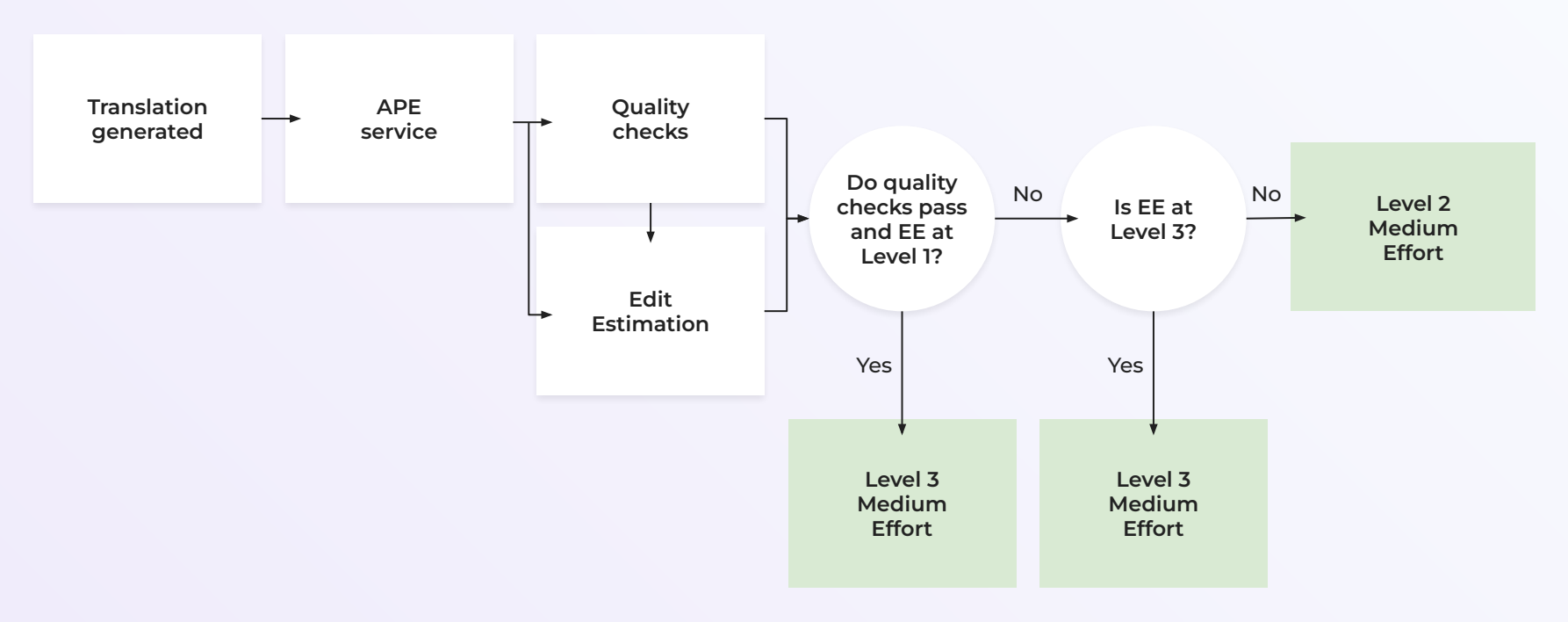
NOTE: Converted to probability.
Lower probability results in higher edit rate.



Post-Processing: Under the hood



Edit Effort Estimation: Under the hood



R&D roadmap

2025 Research Roadmap

RELEASED IN 2024

- AI Translation Features
  - Adaptive translation memory
  - Glossary term insertion with LLMs
  - Informality changes for Spanish, German, and Italian
  - LLMs for translation evaluation
  - Fine-tuning models for more accurate translation evaluations
- Translate directly using GPT

IN RESEARCH AND MVP(Now)

- LLM translation smoothing
- High-performing auto-MQM
- Google PETLLM (LoRA)
- LLMs for higher quality translations
- Choose your favorite LLM
- LLM as MT provider
- Optimize handling of emojis, all caps, hashtags, brackets

UPCOMING (Later)

- Self hosted LLMs
- Expand LLM language support
- Source content pre-editing
- Improvements for translation quality estimation
- Continued work for automating MQM
- LLM content creation
- Prompt Engineering Tools