From Search Relevance to Content Safety

- Research @ *SurreyNLP* -

Dr Diptesh Kanojia

People-Centred AI
UNIVERSITY OF SURREY

NATURE INSPIRED COMPUTING AND ENGINEERING RESEARCH GROUP

Invited Talk | eBay Inc. | 6th June 2025

Introduction / Context

- *SurreyNLP* – Our research group comprising PhD students and researchers @ University of Surrey.
  - Focus on cutting-edge NLP and Computer Vision problems
  - Open, efficient, responsible, and aligned with user needs.

- Presentation today is partly our group's research output across domains relevant to large scale digital platforms.

Talk Outline

- Search Relevance Optimization
  - Challenge: Query Ambiguity & User Intent
  - UCO and NEAR^2
  - Intent and Aspect-based Reasoning
- Language Technology
  - Cross-lingual and Low-resource NLP
  - Quality Estimation and Automatic Post-Editing
  - Translation Error Reasoning and Correction
- Online Safety – Multimodal NLP
  - Hate in Video *vs.* Hate in Memes – Different Challenges?
  - Breaking down the challenges – visually, and with reasoning in-context
  - Efficient Training – The CAMU framework
- Future Directions / Concluding Remarks

Search Relevance Optimization

Research in collaboration with ebay

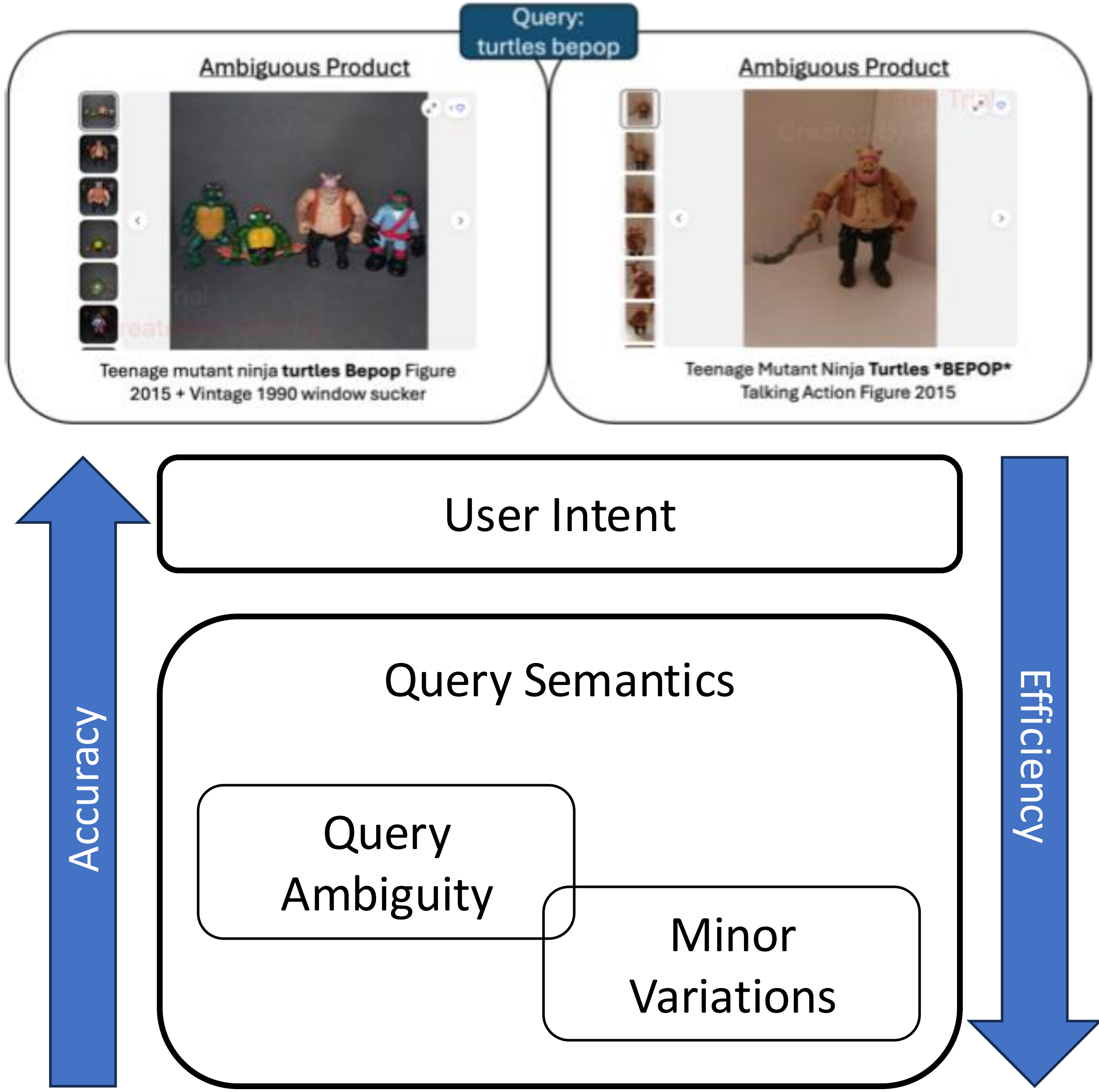
Co-contributors: Samarth Agarwal, Constantin Orasan, Hadeel Saadany, Swapnil Chaudhari, Shenbin Qian, Zhe Wu

Challenges

- <u>Semantic and User Intent mismatches</u> are known problems in the retrieval area.

  e.g., query for product (*iPhone 16*) *vs.* accessory for the product (*iPhone 16* ***cover***), ambiguity in product line (see Figure ->)

- eBERT resolves contextualization issues within semantics **to a certain extent** – more training, more data
  - generalization vs. specificity tradeoff.
- Alphanumeric Queries – problems with minor character variations lead to major differences in products or their aspects.
- Efficiency – Massive real-time product catalog/KG @ eBay – latency issues
  - Accuracy *vs.* Efficiency trade-off

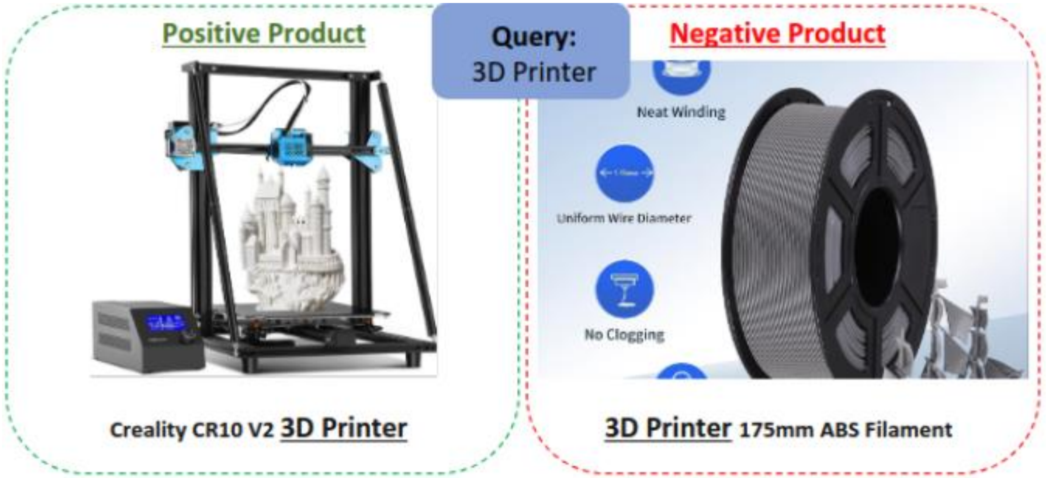


User Intent in Search & Centrality Awareness

To address the intent mismatch, we propose leveraging the concept of **user-intent centrality**.

- **Centrality** is how well a product title centrally matches a user's *expected* result for a query, as opposed to being merely related; **label derived from human annotations.**
- **Hard Negatives** are items that are semantically very similar to positive results but are non-central to the user's intent. For example, "iPhone 13 cover" is a hard negative for the query "iPhone 13".

Test Name	# Corpus	# Queries
CQ	187469	17325
CQ-balanced	46561	17325
CQ-common-str	12508	6351
CQ-alphanum	162115	12333



Central: Thomas sabo charms with 18k Rose gold pearl

Non-central: Thomas Sabo charm club bracelet with detachable dragonfly charm

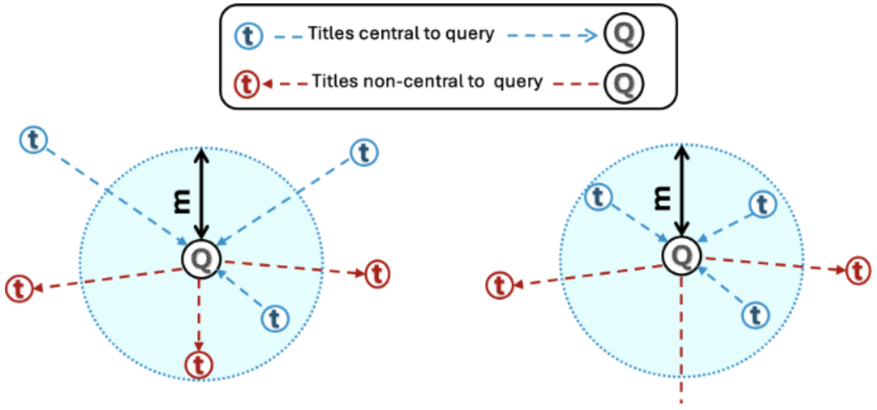
Addressing User Intent: The UCO Approach

- **UCO (User-intent Centrality Optimization)** is a fine-tuning approach for existing encoders (e.g., eBERT) to optimize for the user-intent centrality score
- Fine-tuning performed on an internal eBay dataset with **relevance** and **binary centrality** scores.
- **Dual-Loss Mechanism** A novel combination of two loss functions is used to handle hard negatives:
    - **MNRL (Multiple Negative Ranking Loss) m**inimizes the distance between the query and positive (central) samples while maximizing it for multiple negative samples.

$$\text{MNRL} = \sum_{i=1}^{P} \sum_{j=1}^{N} max(0, f(q, p_i) - f(q, n_j) + margin)$$

    - **OCL (Online Contrastive Loss)** focuses learning on the most challenging pairs (hard positives and hard negatives) within a batch.

$$\text{OCL} = Y * D + (1 - Y) * max(margin - D, 0)^2$$

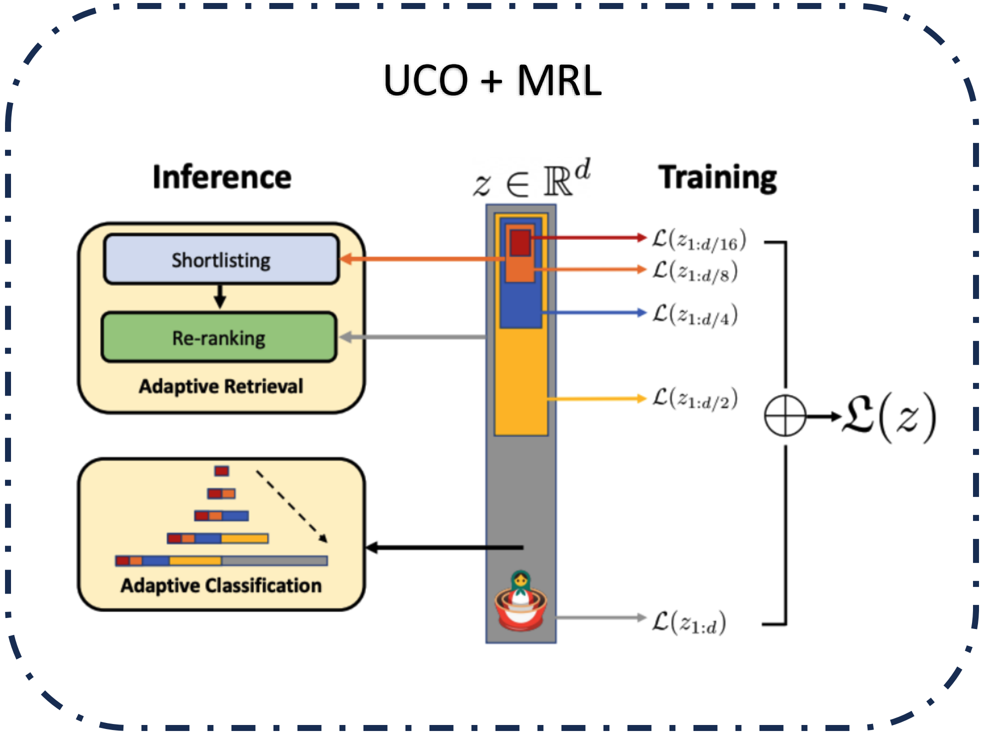


Algorithm targets those non-central titles (red) that are inside the margin

Saadany, H., Bhosale, S., Agrawal, S., Kanojia, D., Orăsan, C., & Wu, Z. (2024). Centrality-aware Product Retrieval and Ranking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Enhancing Efficiency: The NEAR^2 Approach

- **NEAR² (Nested Embedding Approach):** An approach to produce efficient embeddings based on Matryoshka Representation Learning (MRL).
- A single model learns multiple "nested" representations of decreasing dimensionality (e.g., 768, 512, ..., 64) during one training run.
    - Achieved by calculating the task loss for each embedding dimension and taking a weighted sum.
- Eventually, smaller, information-dense embeddings can be used at inference time for significant reductions in model size – this latency!
- Smaller, information-dense embeddings can be used at offline, at inference time.



$$L_{MRL} = \sum_{m \in M} c_m L_{task}(f_m(x), y)$$

Results | Search Optimization

Encoder	UCO	Precision@k (↑) 3	5	10	Recall@k (↑) 3	5	10	NDCG@k (↑) 3	5	10	MRR (↑) @10
***CQ* test**											
BERT	✗	16.20	13.03	8.93	11.31	14.41	18.83	0.1912	0.1818	0.1833	0.2771
eBERT	✗	20.71	17.25	12.54	14.46	19.19	26.26	0.2392	0.2330	0.2430	0.3415
eBERT	✓	64.76	55.74	39.22	49.63	63.92	79.65	0.7439	0.7488	0.7672	0.8189
eBERT (siam)	✗	55.25	48.33	34.90	42.36	56.09	72.22	0.6315	0.6428	0.6704	0.7263
eBERT (siam)	✓	66.25	57.16	40.20	51.18	65.79	81.66	0.7635	0.7698	0.7886	0.8347
***CQ-balanced* test**											
BERT	✗	7.13	4.94	2.95	21.26	24.58	29.33	0.1824	0.1961	0.2115	0.1862
eBERT	✗	9.72	6.94	4.22	29.02	34.58	42.07	0.2428	0.2657	0.2899	0.2495
eBERT	✓	28.57	18.15	9.50	85.40	90.42	94.62	0.7851	0.8059	0.8197	0.7789
eBERT (siam)	✗	25.99	16.68	8.89	77.66	83.08	88.59	0.6888	0.7112	0.7291	0.6784
eBERT (siam)	✓	29.19	18.39	9.58	87.26	91.58	95.43	0.8046	0.8225	0.8351	0.7965
***CQ-common-str* test**											
BERT	✗	9.41	6.31	3.65	28.15	31.47	36.35	0.2532	0.2669	0.2828	0.2579
eBERT	✗	12.62	8.64	5.00	37.79	43.10	49.92	0.3272	0.3491	0.3714	0.3315
eBERT	✓	32.03	19.58	9.92	95.84	97.65	98.87	0.9091	0.9166	0.9206	0.8979
eBERT (siam)	✗	29.93	18.76	9.68	89.57	93.58	96.50	0.8194	0.8361	0.8456	0.8063
eBERT (siam)	✓	32.12	19.64	9.92	96.11	97.94	98.93	0.9117	0.9193	0.9226	0.9003
***CQ-alphanum* test**											
BERT	✗	20.54	16.65	11.47	13.45	17.32	22.82	0.2333	0.2176	0.2226	0.3350
eBERT	✗	23.35	19.54	13.77	15.53	20.76	27.85	0.2630	0.2516	0.2617	0.3739
eBERT	✓	64.58	57.27	40.35	44.05	59.97	77.00	0.7119	0.7094	0.7344	0.8018
eBERT (siam)	✗	60.67	54.10	38.54	41.32	57.10	74.20	0.6652	0.6654	0.6951	0.7618
eBERT (siam)	✓	67.10	59.70	41.81	46.07	62.72	79.76	0.7375	0.7371	0.7609	0.8171

For the CQ test set, eBERT with UCO improved NDCG@10 from 0.2430 to 0.7672

Model	Dimension	Precision@5	Recall@5	NDCG@5	MRR@10
eBERT-siam	768	+13.33%	+11.77%	+13.10%	+10.20%
	512	+13.35%	+11.87%	+13.16%	+10.30%
	256	+13.26%	+11.68%	+13.05%	+10.19%
	128	+13.10%	+11.37%	+12.80%	+10.16%
	64	+11.79%	+9.72%	+11.23%	+9.06%
eBERT-UCO	768	+4.25%	+4.04%	+4.34%	+3.50%
	512	+4.27%	+3.97%	+4.37%	+3.57%
	256	+4.18%	+3.83%	+4.23%	+3.49%
	128	+3.86%	+3.52%	+3.97%	+3.42%
	64	+3.28%	+2.99%	+3.34%	+3.03%
eBERT-siam-UCO	768	+3.85%	+3.75%	+3.82%	+3.05%
	512	+3.85%	+3.72%	+3.81%	+3.00%
	256	+3.62%	+3.47%	+3.61%	+2.96%
	128	+3.46%	+3.27%	+3.46%	+2.96%
	64	+2.75%	+2.45%	+2.77%	+2.58%

Method	eBERT NDCG@5	eBERT MRR@10	eBERT-siam NDCG@5	eBERT-siam MRR@10
MNRL	+4.26%	+3.48%	+2.98%	+2.51%
OCL	+32.09%	+22.50%	+25.86%	+15.66%
MNRL + OCL	+3.34%	+3.03%	+2.77%	+2.58%
MRL: MNRL + OCL	-3.29%	-1.51%	-3.26%	-1.58%

Using a 64-dim embedding (a **12x size reduction**), eBERT-siam with NEAR[2] improved NDCG@5 by **+11.23%** over its 768-dim baseline on the CQ test

Query Intent & Product Aspect | Reasoning | Ongoing Work

Investigating LLM- and Reasoning-based approaches to retrieval of products

- Break down query into its aspects:

Query	Product Title	Media Relevance	Variance	User Intent (Research *vs.* Purchase)	User Intent Explanation	Query Complexity (Simple *vs.* multi-aspect)	Complexity Explanation	Primary Attribute Focus Value	Focus Explanation

- Break down product into aspects:

Product Type (product *vs.* accessory *vs.* Collection)	Exp	Brand Presence	Exp	Complexity Value	Exp	Primary Selling point	Exp	Demographic	Expl	Price Indicator	Product Condition

- Novel Relevance score annotation based on revised guidelines + Centrality Annotation
- GRPO-LoRA Fine-tuning[1]

[1] AlphaMaze: https://arxiv.org/html/2502.14669v3

Insights & Future Directions

- Product retrieval can be substantively improved by concurrently addressing two axes:
  - (1) deeper understanding of user intent (relevance) and,
  - (2) greater computational efficiency (scalability).
- Significant performance improvements using both UCO and NEAR^2 and at a reduced embedding size too!
- **Demonstrated** modeling user intent in IR (UCO) via centrality, and for efficient representation learning for dense retrieval (NEAR^2).
- Reasoning for user-intent and product details may play a key role given LLMs are becoming more accessible.

Future :

A/B testing of these models in production environments to quantify real-world impact.

Investigating Reasoning and GPRO fine-tuning for scalable extension with a pre-hosted LLM.

Extending these techniques to other areas like multimodal search (image-based retrieval).

Investigating unified models that are both intent-aware and computationally efficient by design.

Language Technology Advancement

Research in collaboration with Centre for Translation Studies, University of Surrey; IIT Bombay, India; Tilburg University, Netherlands.

Co-contributors: Constantin Orasan, Fred Blain, Archchana Sindhujan, Sourabh Deoghare, Shenbin Qian, MinnieProf. Pushpak Bhattacharyya

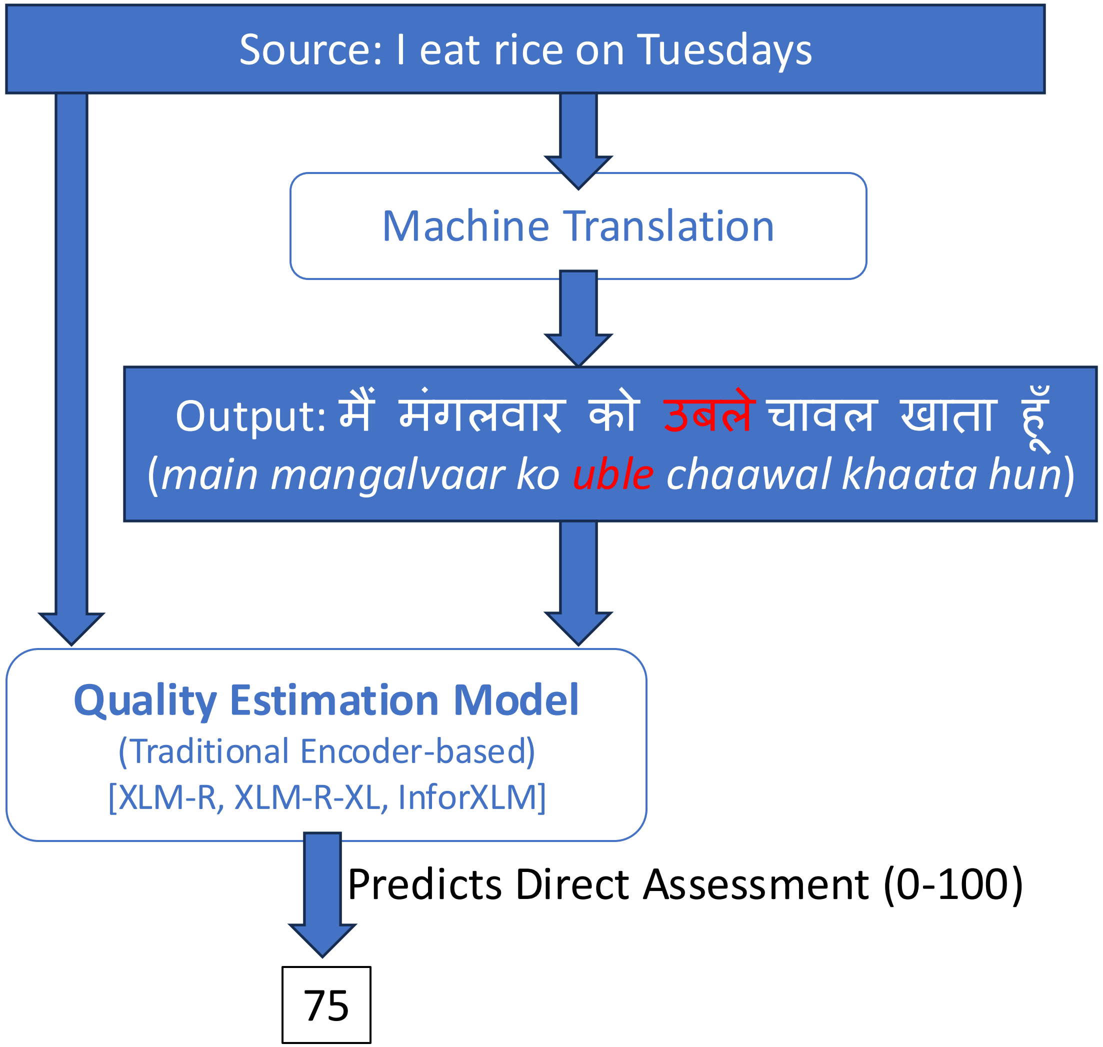
Evaluating Machine Translation - Quality Estimation

To provide a reliable, automatic measure of translation quality, which is crucial for system development and user-facing applications – without using a reference.

- Metrics like BLEU, chrF, MetricX need a reference.
- MT is subjective – multiple references – free order.

Quality Estimation (QE) is task of assessing the quality of machine-translated text in the <u>absence of a human reference translation</u>.

- **Segment-Level QE** focuses on assigning a quality score to a translated sentence, typically a Direct Assessment (DA) score from 0-100.
- **Word-Level QE** focues on tagging each token in source and MT output with a OK/BAD tag, given the translation errors.

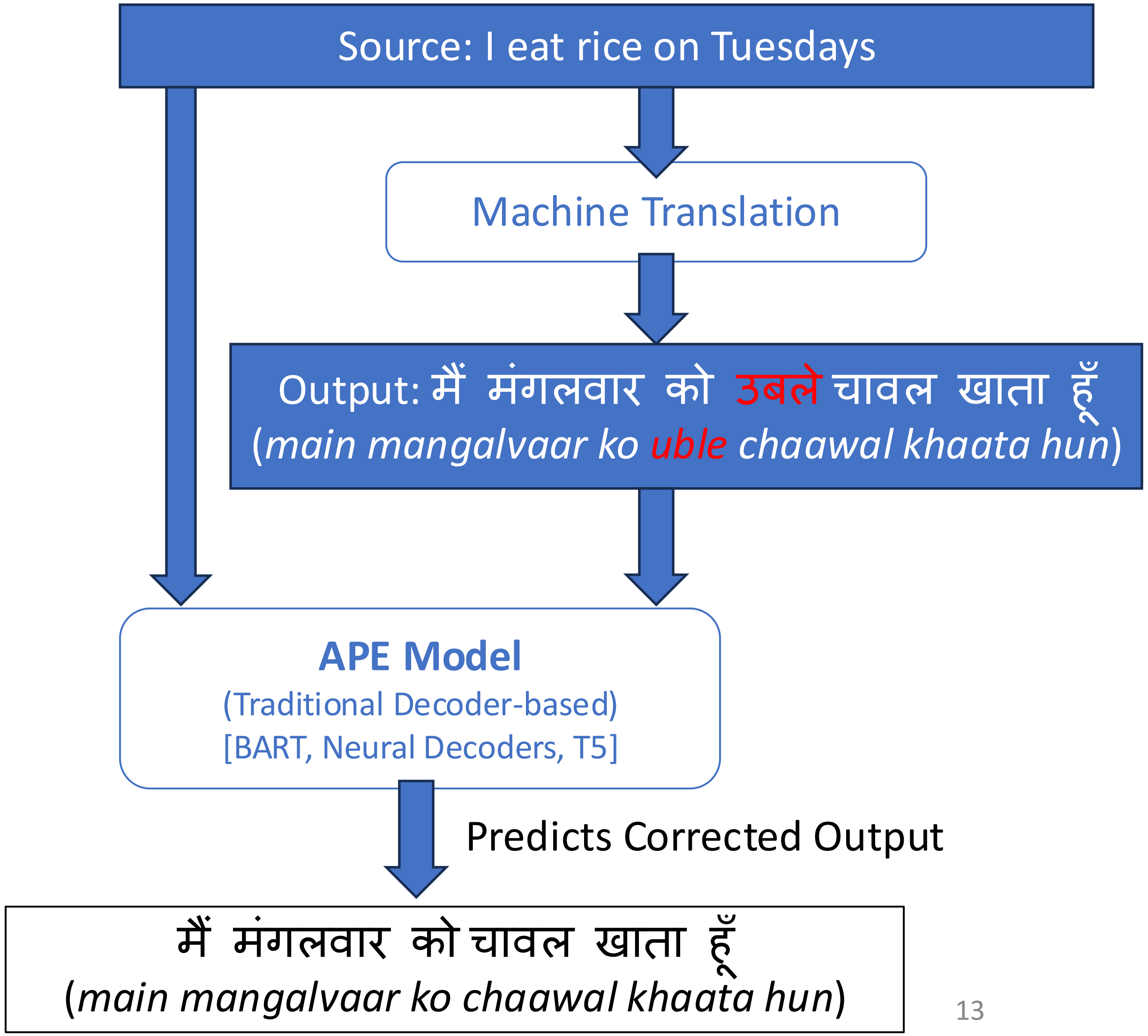


Correcting Machine Translation – Automatic Post-Editing

Particularly valuable in black-box scenarios where the underlying MT model cannot be retrained.

- **Automatic Post-Editing (APE)** is the task of automatically correcting errors in machine-generated translations.

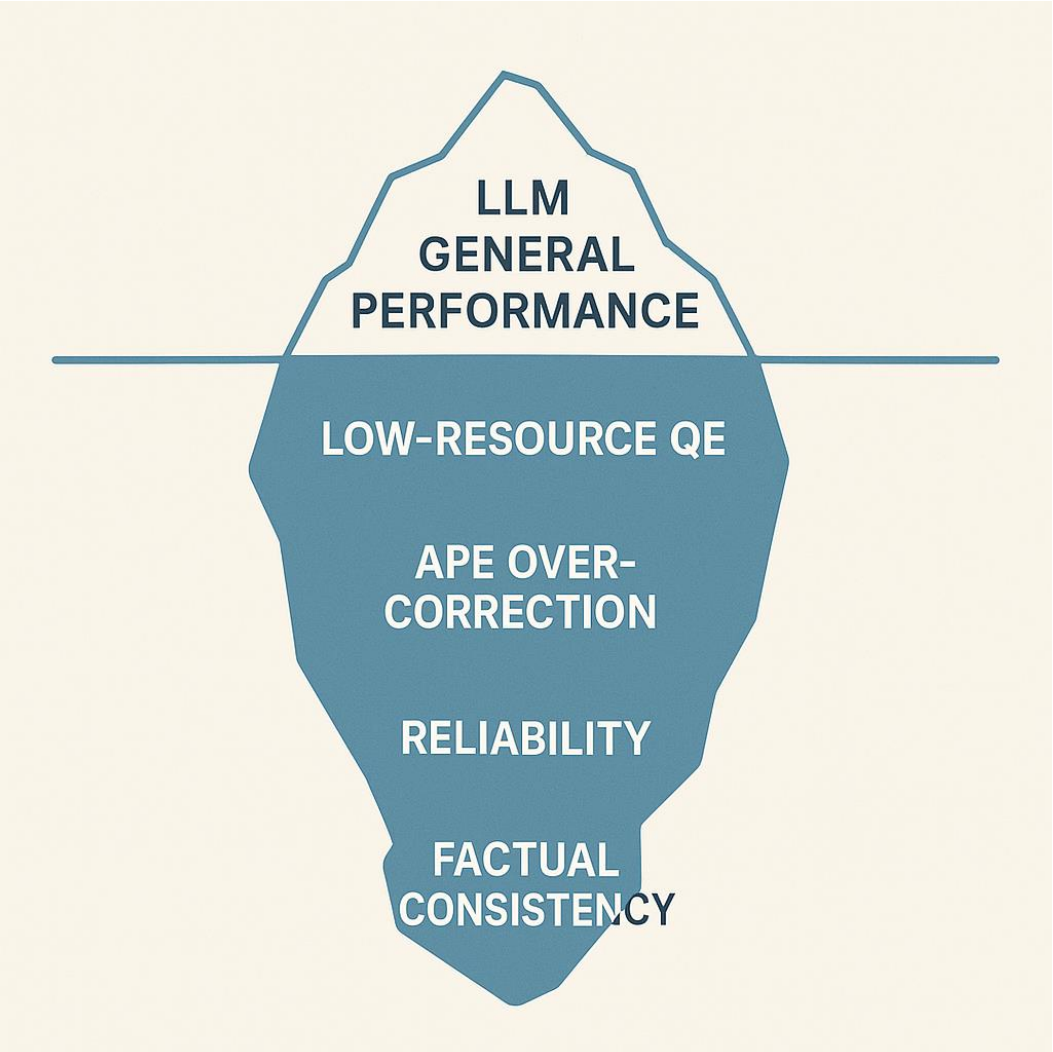
- **Principle of Minimal Editing for data states that** <u>APE systems should aim to make the fewest necessary changes to improve the MT output</u>, preserving fluency and adequacy.



Are we working on a ‘solved’ problem? ;)

- The recent capabilities of Large Language Models (LLMs) have led to claims of superlative performance on many NLP tasks.
- This raises a critical question: Have tasks like QE and APE been effectively "solved" by these large, generalist models?

Our research investigates this assumption, particularly in challenging, real-world scenarios such as reference-less evaluation for low-resource languages.



Enhancing Low-Resource QE – Data & Other Challenges

- No data for Indic languages till 2021
- SurreyNLP contributed to collating the following datasets ->
- Challenges
  - Low-resource languages
  - Long-context, Free word order languages
  - Obscure Languages -> Directionality!
- Cross-linguality -> Comparison across languages
- Domain specificity for Machine Translation?

Lang.	Train	Test
English - Gujarati (En-Gu)	7000	1000
English - Hindi (En-Hi)	7000	1000
English - Marathi (En-Mr)	26 000	699
English - Tamil (En-Ta)	7000	1000
English - Telugu (En-Te)	7000	1000
Estonian - English (Ne-En)	7000	1000
Nepali - English (Ne-En)	7000	1000
Sinhala - English (Si-En)	7000	1000

2025 WMT QE + Metrics Shared Task

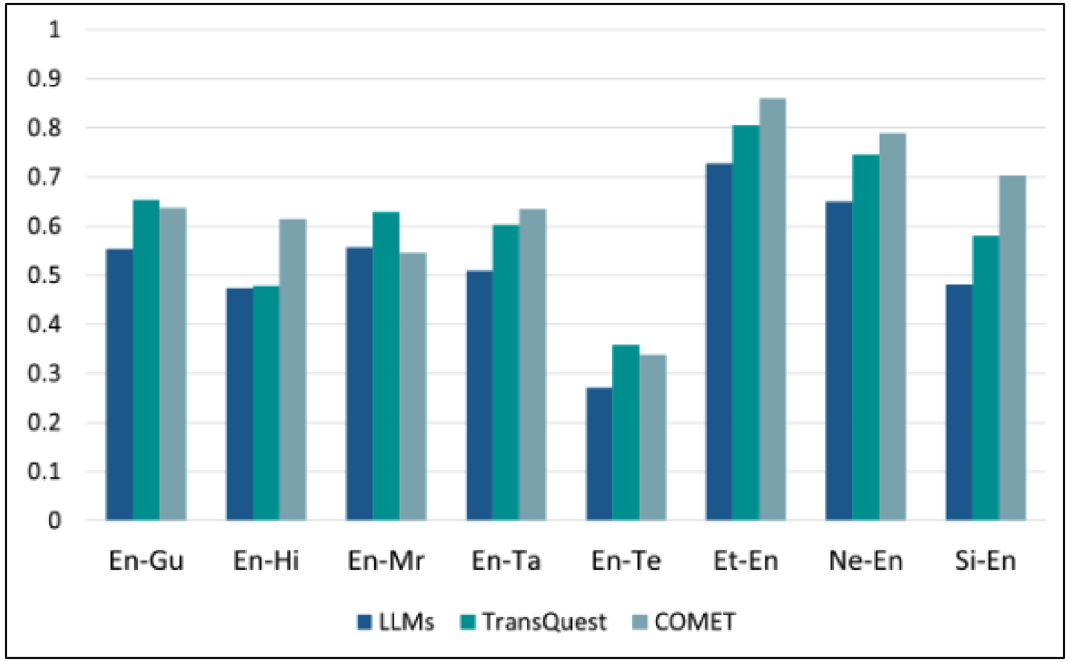
- Much of the research in QE and APE is driven by annual shared tasks at the Conference on Machine Translation (WMT).

- These tasks provide benchmark datasets, standardized evaluation protocols, and a collaborative environment for advancing the state-of-the-art.

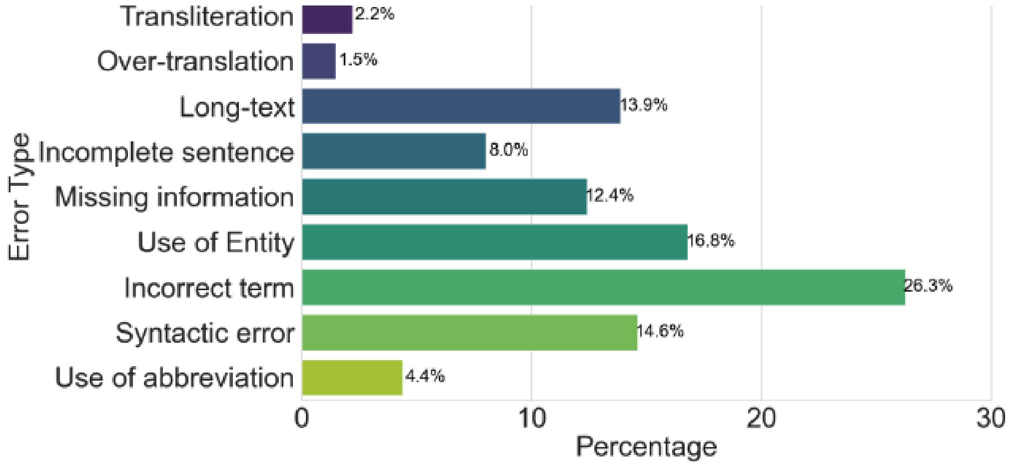
- The datasets used in our QE and APE investigations are primarily from recent WMT shared tasks (*e.g.*, WMT21, WMT22, WMT23, WMT24).

Please see subtask 3 for participating: https://www2.statmt.org/wmt25/mteval-subtask.html#_task_3_quality_informed_segment_level_error_correction

QE-Assisted APE – Bringing it together!



Lang-pair	Gemma-7B	Llama-2-7B	Llama-2-13B	OC-3.5-7B	TransQuest	CometKiwi
Unified Multilingual Training (UMT) Setting						
En-Gu	0.566	0.461	0.465	0.554	0.630	**0.637**
En-Hi	0.449	0.332	0.322	0.458	0.478	**0.615**
En-Mr	0.551†	0.516†	0.505	0.545†	**0.606**	0.546
En-Ta	0.502	0.464	0.471	0.509	0.603	**0.635**
En-Te	0.242	0.258	0.258	0.267	**0.358**	0.338
Et-En	0.728	0.636	0.655	0.678	0.760	**0.860**
Ne-En	0.650	0.519	0.565	0.607	0.718	**0.789**
Si-En	0.455	0.395	0.403†	0.481†	0.579	**0.703**
Independent Language-Pair Training (ILT) Setting						
En-Gu	0.440	0.214	0.421	0.520	**0.653**	-
En-Hi	0.375	0.282	0.336	**0.474**	0.119	-
En-Mr	0.557	0.509†	0.501	0.554†	**0.629**	-
En-Ta	0.475	0.375	0.441	**0.509**	0.303	-
En-Te	0.217	0.263	0.261	**0.271**	0.087	-
Et-En	0.648	0.589	0.598	0.652	**0.806**	-
Ne-En	0.612	0.497	0.543†	0.614	**0.746**	-
Si-En	0.387	0.332	0.346	0.441	**0.581**	-



LLM Performance

LLMs underperform specialized encoder-based models in reference-less QE for low-resource languages, even after instruction fine-tuning.

Tokenization Discrepancy

LLMs tend to over-tokenize morphologically rich, low-resource languages, creating a mismatch with word-level semantics and impacting cross-lingual understanding.

Data Scarcity

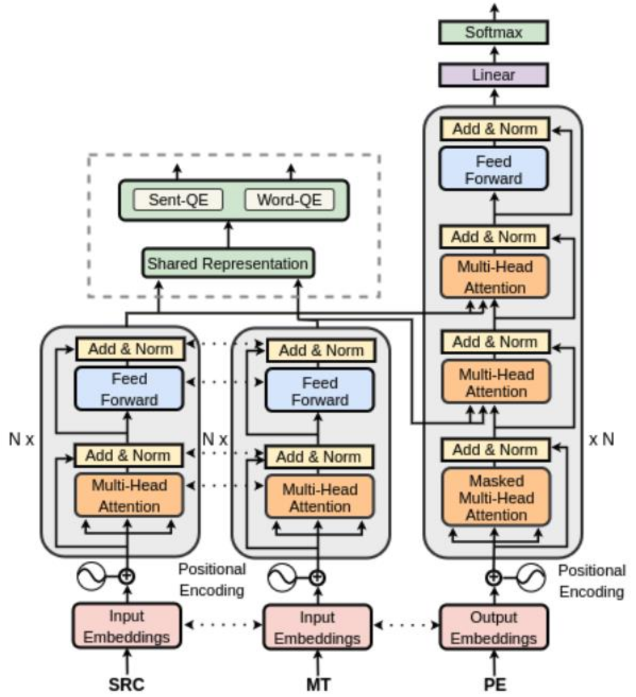
Lack of sufficient annotated data for both QE model training and for pre-training LLMs on these languages.

Scaling our Solutions – Multilingual

- **Hypothesis -** The complementary nature of QE and APE suggests that information from QE can be used to mitigate the "over-correction" problem in APE systems.
  - Propose a **joint multi-task learning (MTL) framework** for QE and APE. [EMNLP 2023]
  - Propose using **word-level QE with Grid-beam Search** at decoding time-step to reduce errors. [NAACL 2025]

- Utilized Nash-MTL, where tasks "bargain" for parameter updates, to jointly train a single model on sentence-level QE, word-level QE, and APE.

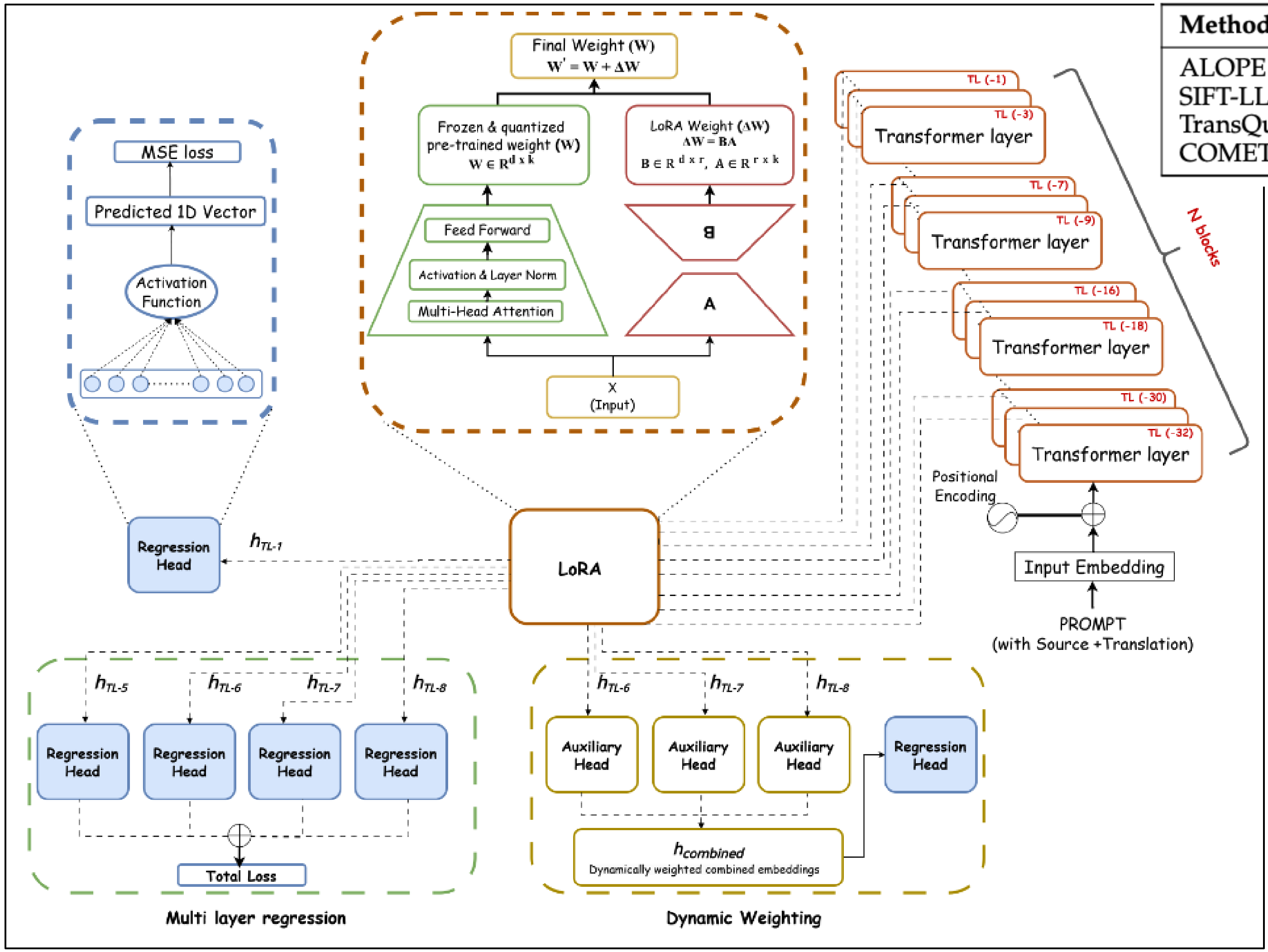
This **tight coupling of QE and APE proved superior to pipeline-based strategies**, reducing over-correction and improving APE performance (+1.09 TER for En-Mr).



Experiment	En-De	En-Hi	En-Mr
Do Nothing	19.06	47.43	22.93
Standalone-APE + BS	18.91	21.48	19.39
QE-APE + BS	18.45	19.75	18.30
Standalone-APE + GBS	18.26	19.62	17.95
QE-APE + GBS	**18.04**	**19.20**	**17.53**
Standalone-APE + GBS (Oracle)	17.74	19.43	17.31
QE-APE + GBS (Oracle)	17.50	18.52	16.70
Greedy	19.38	20.04	18.73
Sampling	19.35	19.89	18.46
top-k Sampling	18.43	19.46	18.18
Lopes et al. (2019)	18.38	19.41	18.16
Deguchi et al. (2024)	18.40	19.93	18.92

Scaling our Solutions – LLM-based QE and APE

Adaptive Layer Optimization for Translation Quality Estimation using Large Language Models (ALOPE)



Method	En-Gu	En-Hi	En-Mr	En-Ta	En-Te	Et-En	Ne-En	Si-En
ALOPE	0.606	0.479	**0.636**	0.610	**0.388**	0.751	0.682	0.573
SIFT-LLMs	0.555	0.393	0.530	0.586	0.290	0.728	0.618	0.558
TransQuest	0.630	0.478	0.606	0.603	0.358	0.760	0.718	0.579
COMET	**0.637**	**0.615**	0.546	**0.635**	0.338	**0.860**	**0.789**	**0.703**

ALOPE proposes using n-embedding layers of an LLM to extract embeddings

Fine-tunes a LoRA with proposed regression head for more deterministic predictions.

Break SoTA barrier for LLM-based QE and beats COMET for 2 language pairs.

Interesting Insights:

Embedding LLM layers 7 to 11 show superlative performance at cross-lingual tasks

Similar observation for multilingual tasks for other languages

Are we over tuning the final layer to task specificity post-training?

The **memory consumption** of our ALOPE-based models is approximately

LLaMA3.2-3B: ~12.8 GB
LLaMA3.1-8B: ~12.7 GB
LLaMA2-7B: ~14.4 GB
Aya-expanse-8B: ~11.9 GB

In comparison, encoder-based SOTA models consume:
TransQuest (InfoXLM): ~11.9 GB
COMET (XLM-R XL): ~ 15 GB

Insights and Future Directions

- For low-resource language tasks like QE, specialized models that leverage linguistic knowledge (e.g., language relatedness) often outperform larger, general-purpose LLMs.
- The synergy between QE and APE is best realized through tight integration, such as joint multi-task learning, which effectively addresses practical issues like over-correction.
- Test-time approaches to APE can be informed by QE
- LLM-based approaches can support MT, QE, and APE - all at once.
- **Future Directions**
  - Improving LLM robustness for cross-lingual tasks via better tokenization
  - Developing unified native cross-lingual QE-APE models, LLM adapters
  - Exploring test-time decoding constraints and contextualization further.

Multimodal NLP

Co-contributors: Girish Kaushik, Helen Treharne, Zhenhua Feng, Muhammad Awais Rana, Aditya Joshi

Online Safety – Multimodal Challenges

- The proliferation of hateful content on social media has moved beyond text to include images, videos, and memes.
  - This necessitates effective detection methods that can analyze content across different modalities (textual, auditory, visual).
- Existing research has focused on unimodal hate speech detection.
  - <u>Effectiveness of approaches</u> across different modality combinations <u>was not well understood</u>.
- Community moderation is not a scalable approach.
- Detecting hate in multimodal content becomes necessary
  - Child-content being targeted with hidden hateful audio. [on Youtube; as of 27th May 2025]

Model	Embeddings	Dataset
BERT (bert-base-uncased)	Text	HatefulMemes, HateMM
HateXplain	Text	HatefulMemes, HateMM
CLIP (clip-vit-base-patch32)	Image, Text	HatefulMemes, HateMM
ViT (vit-base-patch16-224-in21k)	Image	HatefulMemes, HateMM
DINOv2 (dinov2-small)	Image	HatefulMemes, HateMM
CLAP (clap-htsat-unfused)	Audio, Text	HateMM
MFCC	Audio	HateMM
AudioVGG19	Audio	HateMM
Wav2Vec2 (wav2vec2-base-960h)	Audio	HateMM

Table 1: Encoder models for different modalities

Contextualizing Hate: Challenges

Nuanced Cross-Modal Interactions

- Hatefulness in memes often arises from the complex interplay between visual and textual cues, not from either modality in isolation.

Benign Confounders

- Challenge where a hateful meme can be made non-hateful (or vice-versa) by changing only the image or the text; makes it difficult for models (fusion-based approaches) to learn true cross-modal understanding.

Cultural Context

- Detecting hateful memes requires an understanding of underlying linguistic and cultural contexts that distinguish hateful rhetoric from benign humour.

Data Annotation

- Annotation Bias for hateful commentary given political stance.

Fusion Approaches – Limited Applicability

	Models	F1 (M)	P (M)	R (M)	Acc
Existing	BERT + MFCC + ViT [12]	0.749	0.742	0.758	0.798
	HXP + MFCC + ViT [12]	0.720	0.718	0.726	0.777
	BERT + AVGG19 + ViT [12]	0.718	0.723	0.719	0.755
	HXP + AVGG19 + ViT [12]	0.707	0.714	0.712	0.767
Sim Fusion	HXP + CLAP + ViT (Concat)	0.823	0.803	0.765	0.832
	CLAP Text + CLAP Audio + CLIP (Concat)	0.802	0.788	0.741	0.811
	HXP + CLAP + CLIP (Concat) **(HCC1)**	**0.848**	**0.840**	0.800	**0.854**
MO-Hate	BART → Wav2Vec2 → DINOv2	0.821	0.822	**0.820**	0.820
	Wav2Vec2 → ViT → Text (BART)	0.794	0.794	0.794	0.794
	BART → CLAP → DINOv2 **(BCD1)**	0.821	0.821	0.820	0.820

	Models	AUROC	F1 (M)	P (M)	R (M)	Acc
0-shot VLMs	OpenFlamingo (7B) [3]	0.570	-	-	-	0.564
	LLaVA-1.5 (13B) [28]	0.618	-	-	-	0.614
	InstructBLIP (13B) [29]	0.596	-	-	-	0.601
	Evolver (13B) [24]	0.603	-	-	-	0.604
Existing	**PALI-X-VPD** [23]	**0.892**	-	-	-	-
	RGCL-HateCLIPper [32]	0.867	-	-	-	0.788
	Flamingo - fine-tuned [1]	0.866	-	-	-	-
	Hate-CLIPper - Align [26]	0.858	-	-	-	-
Sim Fusion	HXP + CLIP (Concat)	0.615	0.557	0.643	0.492	0.617
	CLIP Text + CLIP Image (Concat)	0.606	0.531	0.644	0.451	0.609
	CLIP Text + CLIP Image (EW Product)	0.591	0.467	0.660	0.361	0.596
MO-H	BART → CLIP	0.618	0.608	0.637	0.622	0.622
	BART → DINOv2 **(MBD1)**	0.628	0.619	0.645	0.631	0.631

- The effectiveness of fusion architectures is highly dependent on the modality combination.
- Simple embedding fusion achieves state-of-the-art performance on video content (HateMM dataset), with a 9.9% F1-score improvement over baseline.
- However, these same fusion approaches fail to capture the nuanced semantic relationships in image-text memes, performing poorly on the HMC dataset.

Paper won the best paper award at MM4SG workshop at WWW 2025, Sydney, Australia; last month!

Breaking the challenge visually

To design a framework that can handle the nuanced challenges of hateful memes where simple fusion fails.

- Instead of just fusing existing representations, we can augment the context available to the model before classification.
- The approach should break down visual components and text on image too?
- Should it be able to segment and identify objects within, too?
- Would the overall approach be efficient?

Disclaimer: Hateful Meme shown for research purposes; only to be demonstrative

(a) True Label: Not Hateful, Prediction: Not Hateful

(b) True Label: Hateful, Prediction: Not Hateful

CAMU: Context Augmentation for Meme Understanding

CAMU proposes

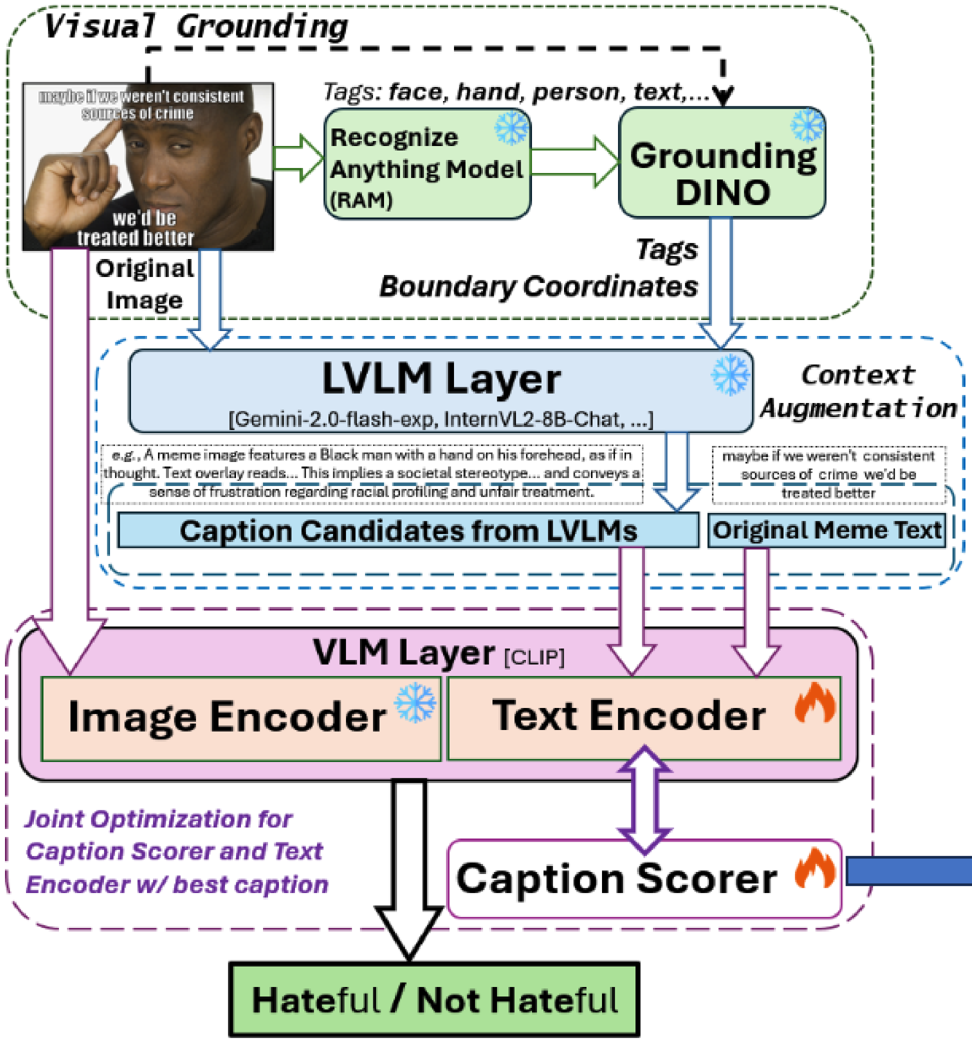
- Visual Grounding
- Context Augmentation
- Efficient fine-tuning
- Joint optimization

VG uses object detection models (RAM, GroundingDINO) to identify key visual elements in the meme image.

CA Leverages Large Vision-Language Models (LVLMs) to generate descriptive captions that incorporate the grounded visual details and original text.

Joint Optimization of the novel caption scorer selecting the most relevant caption, which is then used for parameter-efficient fine-tuning (PEFT) of only the deeper layers of CLIP's text encoder for final classification.

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{cls}}}_{\text{classification}} + \lambda_1 \underbrace{\mathcal{L}_{\text{rel}}}_{\text{hate alignment}} + \lambda_2 \underbrace{\mathcal{L}_{\text{cont}}}_{\text{contrastive}}$$



Caption Scorer, S_i,

$$S_i = f_\theta(\mathbf{h}_i) = \mathbf{W}_5(\phi_4(\mathbf{W}_4(\phi_3(\mathbf{W}_3(\phi_2(\mathbf{W}_2(\phi_1(\mathbf{W}_1\mathbf{h}_i))))))))$$

where:

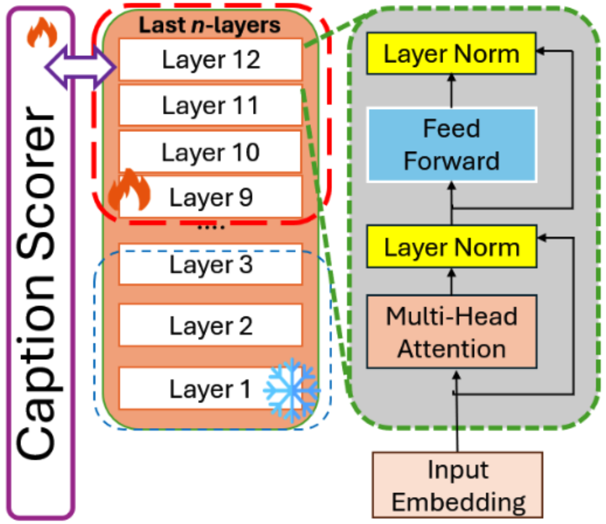
$$\phi(x) = \text{GELU}(\text{LayerNorm}(x)) \odot \text{Dropout}(p)$$

$$\mathbf{W}_1 \in \mathbb{R}^{d\times 1024}, \quad \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{1024\times 1024}, \mathbf{W}_4 \in \mathbb{R}^{1024\times 512}, \quad \mathbf{W}_5 \in \mathbb{R}^{512\times 1}$$

Gumbel-softmax

(differentiable caption selection)

$$p_i = \frac{exp((s_i + g_i)/\tau}{\Sigma_{j=1}^{n} exp((s_j + g_j)/\tau}$$



Disclaimer: Hateful Meme shown for research purposes; only to be demonstrative

Results and Insights

Model	Fine-tuning	AUROC	Acc.	F1	P	R
CLIP-ViT-B/16	Text encoder	0.788	0.632	0.591	0.701	0.626
CLIP-ViT-L/14 (w/ $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{cont}}$)	Text (Last Layer)	0.787	0.717	0.716	0.724	0.718
	Text (Last 2)	0.801	0.712	0.702	0.736	0.708
	Text (Last 3)	0.808	0.736	0.735	0.744	0.737
	Text (Last 4)	0.812	0.753	0.752	0.759	0.754
CLIP-RoBERTa-ViT	Text encoder	0.704	0.621	0.476	0.738	0.351
CLIP-XLM-R-ViT-H/14 (w/ $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{cont}}$)	Text (Last Layer)	0.795	0.730	0.725	0.755	0.733
	Text (Last 2)	0.807	0.742	0.739	0.758	0.744
	Text (Last 3)	0.791	0.747	0.746	0.753	0.748
	Text (Last 4)	0.819	<u>0.775</u>	<u>0.774</u>	<u>0.783</u>	<u>0.776</u>
CLIP-ViT-L/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{cont}$	Text (Last 4)	0.806	0.715	0.708	0.742	0.718
CLIP-ViT-L/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{rel}$	Text (Last 4)	<u>0.824</u>	0.755	0.753	0.765	0.757
CLIP-XLM-R-ViT-H/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{cont}$	Text (Last 4)	0.803	0.754	0.751	0.773	0.756
CLIP-XLM-R-ViT-H/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{rel}$	Text (Last 4)	**0.849**	**0.807**	**0.806**	**0.813**	**0.808**

Fine-tuning	Model	AUROC	Accuracy	F1	Precision	Recall
Text (last layer)	CLIP-ViT-L/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{cont}$	0.809	0.737	0.737	0.740	0.738
	CLIP-ViT-L/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{rel}$	<u>0.829</u>	<u>0.743</u>	<u>0.740</u>	<u>0.757</u>	<u>0.745</u>
	CLIP-XLM-R-ViT-H/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{cont}$	0.772	0.727	0.722	0.751	0.730
	CLIP-XLM-R-ViT-H/14 w/ $\mathcal{L}_{cls} + \mathcal{L}_{rel}$	**0.834**	**0.774**	**0.771**	**0.794**	**0.776**

- CAMU framework achieves high accuracy (0.807) and F1-score (0.806) on the Hateful Memes dataset, performing on par with much larger SoTA models (55B params) while being significantly more efficient.
- For complex multimodal tasks like meme analysis, explicitly augmenting context and using targeted, parameter-efficient fine-tuning is more effective than simple fusion of pre-computed embeddings.

Future Directions | Content Generation

Future Research in Detection: Development of unified frameworks that incorporate modality-specific architectural considerations.

- Improving visual grounding to capture subtle objects or context currently missed by detectors.
- Exploring the use of intermediate-layer representations from encoders to capture distinct semantic nuances.

Beyond this -> Content Generation: Investigating the generation of "counter-narratives" or explanations for why a piece of content is flagged.

- Leveraging generative models to create challenging new test cases (e.g., novel benign confounders) to build more robust detection systems.
- Dance Generation: PhD Student: Xinran Liu; Diffusion-based Music-driven Dance Generation
- Audio-to-Talking Face: PhD Student: Fatemeh Nazarieh; Transformers + CNN, and Diffusion-based approaches to audio-driven talking face generation.

Summary

- **E-commerce Search Optimization**
  - We presented two approaches to enhance search – While UCO method improves relevance by modeling user intent/centrality, NEAR² significantly improves efficiency nested embeddings
  - Enabling up to a 12x reduction in model size with comparable accuracy.
- **Language Technology Advancement**
  - Identified limitations of LLMs in reference-less, low-resource scenarios due to factors like tokenization, linguistic issues.
  - We showed that leveraging language relatedness can improve specialized QE models.
  - Subsequently, we demonstrated that jointly training QE and Automatic Post-Editing (APE) systems via multi-task learning effectively mitigates the problem of over-correction.
- **Multimodal NLP for Online Safety**
  - Analyzed the challenges of multimodal hate detection, finding that fusion approaches effective for video content are not sufficient for complex image-text memes.
  - To address this, we presented the CAMU framework, an efficient method that uses context augmentation and targeted fine-tuning to achieve performance on par with much larger models for hateful meme detection.

Concluding Remarks

- Specialized vs. Generalist models
- Critical Role of Data and Context
- Synergy in AI tasks (inclusive of languages and modalities)
- Towards Scalable and Efficient AI

Thank you!

- Questions?
- Contact: d.kanojia@surrey.ac.uk

- SurreyNLP Github: https://github.com/surrey-nlp
- SurreyNLP Huggingface: https://huggingface.co/surrey-nlp

Acknowledgements

People-Cented AI, NICE, Centre for Translation Studies, Zibanka Media & Techliebe (annotation agencies for QE/APE data), eBay, EAMT.

References

- **Search Relevance Optimization**
  - Saadany, H., Bhosale, S., Agrawal, S., Kanojia, D., Orăsan, C., & Wu, Z. (2024). Centrality-aware Product Retrieval and Ranking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*.
  - Qian, S., Kanojia, D., Agrawal, S., Saadany, H., Bhosale, S., Orasan, C., & Wu, Z. (2025). NEAR²: A Nested Embedding Approach to Efficient Product Retrieval and Ranking. In *SIGIR Workshop on eCommerce 2025.*
- **Language Technology Advancement**
  - Qian, S., Sindhujan, A., Kabra, M., Kanojia, D., Orăsan, C., Ranasinghe, T., & Blain, F. (2024). What do Large Language Models Need for Machine Translation Evaluation?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
  - Sindhujan, A., Kanojia, D., Orăsan, C., & Qian, S. (2025). When LLMs Struggle: Reference-less Translation Evaluation for Low-resource Languages. *arXiv preprint arXiv:2501.04473*.
  - Sindhujan, A., Kanojia, D., & Orăsan, C. (2024). Optimizing Quality Estimation for Low-Resource Language Translations: Exploring the Role of Language Relatedness. In *Proceedings of New Trends in Translation and Technology (NeTTT)*.
  - Deoghare, S., Kanojia, D., Ranasinghe, T., Blain, F., & Bhattacharyya, P. (2023). Quality Estimation-Assisted Automatic Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
  - Deoghare, S., Kanojia, D., & Bhattacharyya, P. (2025). Giving the Old a Fresh Spin: Quality Estimation-Assisted Constrained Decoding for Automatic Post-Editing. *arXiv preprint arXiv:2501.17265*.
  - Deoghare, S., Kanojia, D., & Bhattacharyya, P. (2024). Together We Can: Multilingual Automatic Post-Editing for Low-Resource Languages. *arXiv preprint arXiv:2410.17973*.
- **Online Safety - Multimodal NLP**
  - Koushik, G. A., Kanojia, D., & Treharne, H. (2025). Towards a Robust Framework for Multimodal Hate Detection: A Study on Video vs. Image-based Content. *arXiv preprint arXiv:2502.07138*.
  - Koushik, G. A., Kanojia, D., Treharne, H., & Joshi, A. (2025). CAMU: Context Augmentation for Meme Understanding. *arXiv preprint arXiv:2504.17902*.