– WNNLP 2025 –
The 7th IN5550 Teaching Workshop on Neural Natural Language Processing

University of Oslo

June 2nd 2025



WNNLP 2025 at a glance

- 19 accepted papers submitted by 35 students (30 MSc + 5 PhDs)
- 7 different thematic tracks:
  1. Full dependency parsing of Norwegian (Track Chair: David Samuel)
  2. Parameter Efficient Definition Modeling (Track Chair: Mariia Fedorova)
  3. Food hazard detection (Track chair: Sondre Wold)
  4. Pretraining Tiny Language Models (Track Chair: Lucas Georges Gabriel Charpentier)
  5. Preference prediction and explanation (Track Chair: Vladislav Mikhailov)
  6. Fact verification (Track Chair: Roxana Pop)
  7. Unlearning sensitive content from LLMs (Track Chair: Andrey Kutuzov)
- All papers peer-reviewed by at least two students + one track chair.
- Workshop proceedings: to be published soon

Program

- 14:15 - 14:20 Welcome and awards
- 14:20 - 14:22 Food hazard detection (FHD) task overview
  - 14:22 - 14:41 5 FHD lightning talks
- 14:41 - 14:43 Definition modeling (DEM) task overview
  - 14:43 - 14:46 1 DEM lightning talk
- 14:46 - 14:48 Training tiny LLMs (TLM) task overview
  - 14:48 - 14:54 2 TLM lightning talks
- 14:54 - 14:56 Preference prediction (PPE) task overview
  - 14:56 - 15:00 1 PPE lightning talk
- 15:00 - 15:15 Break
- 15:15 - 15:17 Unlearning content from LLMs (ULM) task overview
  - 15:17 - 15:36 5 ULM lightning talks
- 15:36 - 15:38 Dependency parsing of Norwegian (DPN) task overview
  - 15:38 - 15:44 2 DPN lightning talks
- 15:44 - 15:46 Fact verification (FAV) task overview
  - 15:46 - 15:57 3 FAV lightning talks
- 15:57 - 16:00 Closing

Awards

- Only MSc students considered for awards, not PhD fellows.

Awards

- Only MSc students considered for awards, not PhD fellows.

Best paper

Awards

- Only MSc students considered for awards, not PhD fellows.

Best paper

- Authors: Aliaan Nadeem Azam, Soban Ali, Sander Harald Sævik
- Title: Balancing Accuracy, Efficiency, and Training Cost in Runtime-Constrained Food Hazard Classification

Awards

- Only MSc students considered for awards, not PhD fellows.

Best paper

- Authors: Aliaan Nadeem Azam, Soban Ali, Sander Harald Sævik
- Title: Balancing Accuracy, Efficiency, and Training Cost in Runtime-Constrained Food Hazard Classification

Highest scores across obligatory assignments

Awards

- Only MSc students considered for awards, not PhD fellows.

Best paper

- Authors: Aliaan Nadeem Azam, Soban Ali, Sander Harald Sævik
- Title: Balancing Accuracy, Efficiency, and Training Cost in Runtime-Constrained Food Hazard Classification

Highest scores across obligatory assignments

- Nora Winger Heggelund and Eivind Morris Bakke, with full 32 points both
- As a team, topped the obligatory 2 leaderboard.
- Great job!

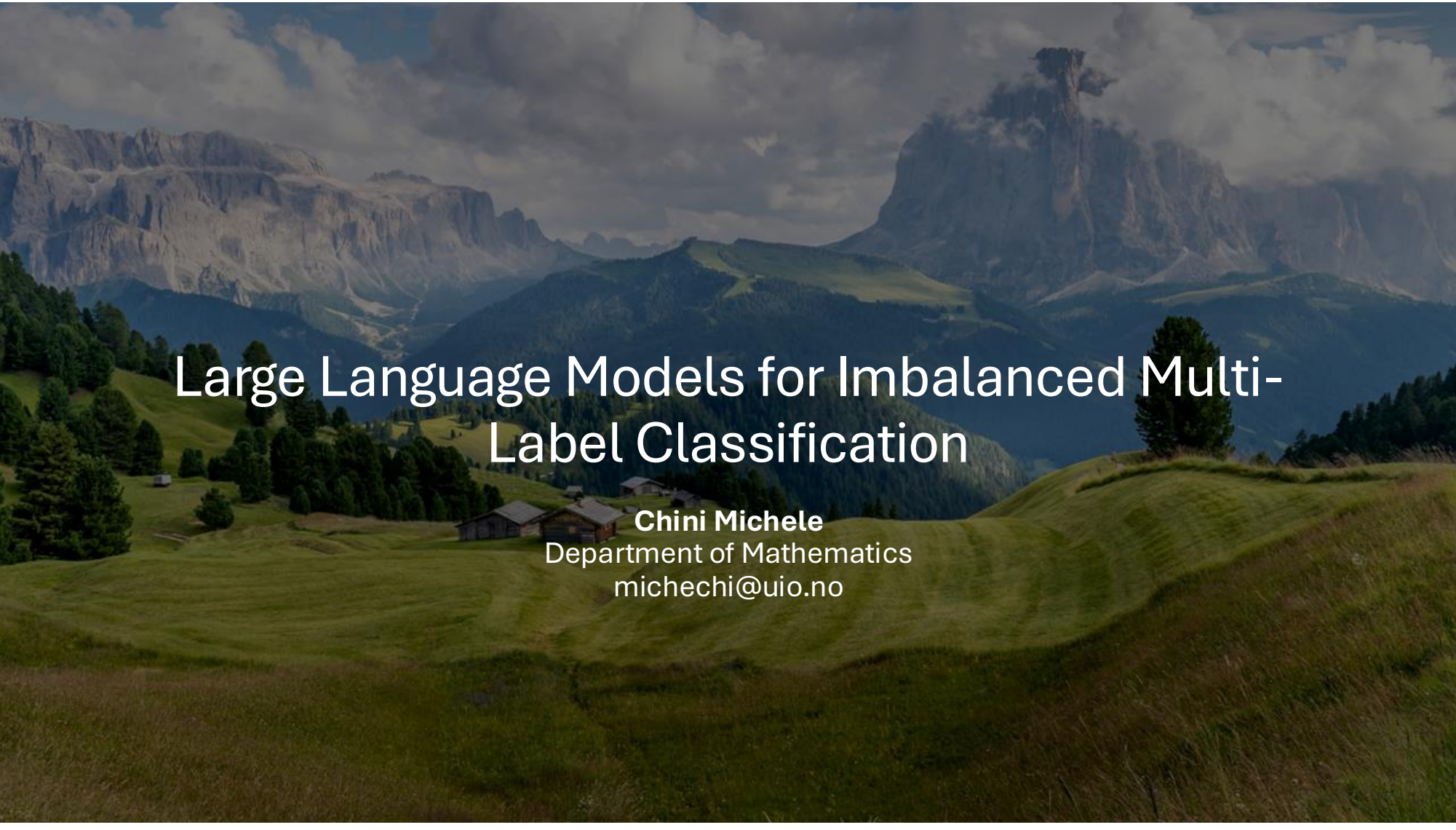
Task overview: Food hazard detection

Goal

- Food contamination is bad.
- Governmental food agencies have to monitor thousands of food recall statements.
- Making a classifier that categorizes food recall statements according to their hazard and product catecory.
- The exam track is based on the Food Hazard Detection Challenge from SemEval 2025.

"Randsland brand Super Salad Kit recalled due to Listeria monocytogenes"	
hazard:	listeria monocytogenes
hazard-category:	biological
product:	salads
product-category:	fruits and vegetables

"Create Common Good Recalls Jambalaya Products Due To Misbranding and Undeclared Allergens"	
hazard:	milk and products thereof
hazard-category:	allergens
product:	meat preparations
product-category:	meat, egg and dairy products



Large Language Models for Imbalanced Multi-Label Classification

Chini Michele
Department of Mathematics
michechi@uio.no

Research question

- How can we improve multi-class, multi-label classification of food hazard texts in extremely imbalanced datasets?
- Many classes are severely underrepresented.
- Hard to accurately classify rare classes in such data.

Models & Pre-processing

- **Baseline Models**:
  - **Logistic Regression** and **XGBoost** with TF-IDF features.
  - Encoder-only transformer (**RoBERTa**) with dual-classification heads.
- **Pre-processing**:
  - Used *title* + *text* features (both are the best).
  - Summarization of *text* with T5-base (didn't improve performance).

Data Augmentation Strategies

- **Oversampling**: Increased minority class samples by duplicating data.
- **Back-translation**: Created text variations via English ↔ French translations using **NLLB-200-distilled-600M**.
- **Few-shot generation**:
  - Used **LLaMA-2 7B** to generate synthetic examples.
  - 3-shots approach.

Key Findings & Takeaways

- **LLMs (RoBERTa)** consistently outperformed traditional models, especially when using both title and text.
- **Summarization** didn’t help; likely removed useful context.
- **Few-shot data augmentation** with LLaMA-2 7B achieved the best performance (conditional F1 0.78 for ST1 and 0.46 for ST2).
- **Overly aggressive augmentation** hurt performance
- **Future work**: Better quality control for generated data, explore other transformer architectures, and tune hyperparameters more systematically.

References

1. Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. Learning from imbalanced data sets, volume 10. Springer.
2. Mihalis Papakonstantinou, Manos Karvounis, Giannis Stoitsis, and Nikos Manouselis. 2022. Deploying a Scalable Big Data Platform to Enable a Food Safety Data Space, pages 227–248. Springer International Publishing, Cham.
3. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limitsof transfer learning with a unified text-to-text transformer. Preprint, arXiv:1910.10683.
4. Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. CICLe: Conformal in-context learning for largescale multi-class food risk classification. In Findings of the Association for Computational Linguistics: ACL 2024, pages 7695–7715, Bangkok, Thailand. Association for Computational Linguistics.
5. Manon Reusens, Alexander Stevens, Jonathan Tonglet, Johannes De Smedt, Wouter Verbeke, Seppe vanden Broucke, and Bart Baesens. 2024. Evaluating text classification: A benchmark study. Expert Systems with Applications, 254:124302.
6. Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Assessing the Impact of Data Augmentation on Food Hazard Detection



WNLLP'2025
02.06.2025

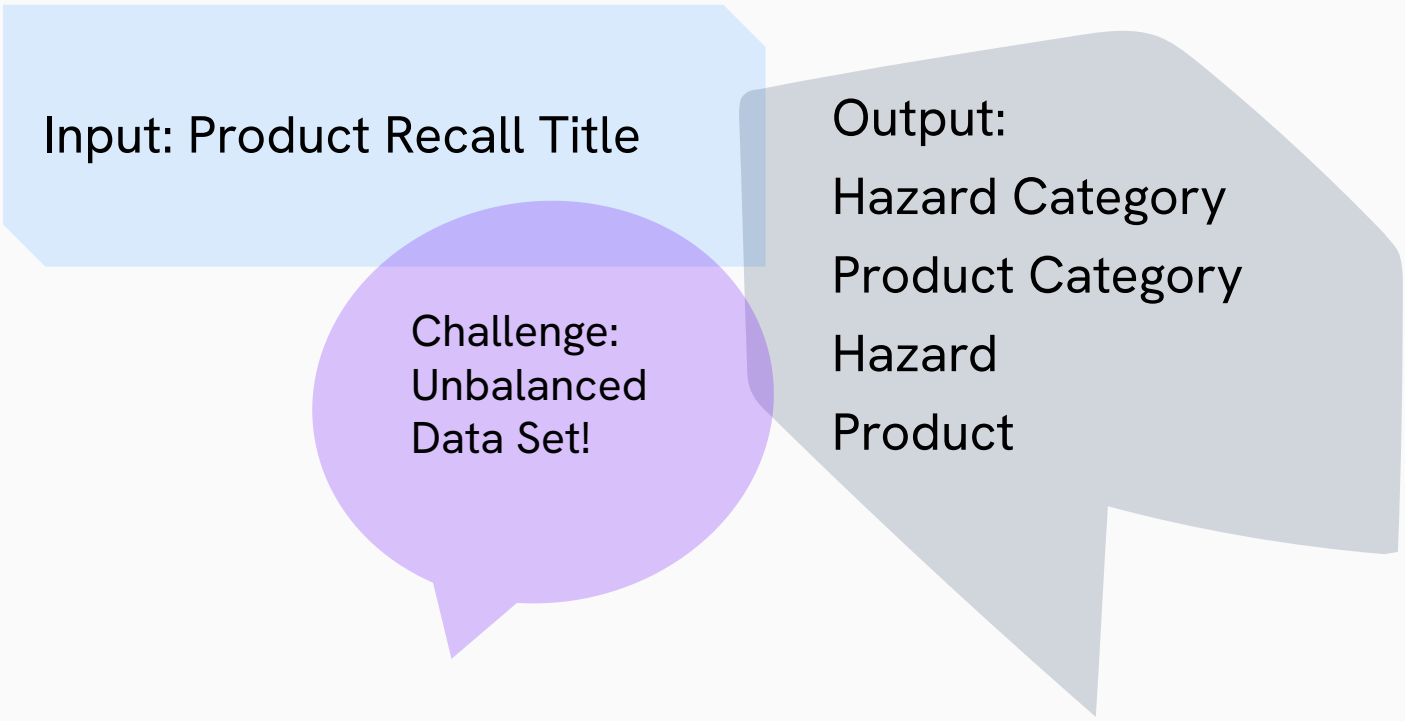
Mina Sheikhi
Iris S. Bore
Elisabeth Othamar

Recap: Task

Input: Product Recall Title

Output:
Hazard Category
Product Category
Hazard
Product

Recap: Task


Input: Product Recall Title
Challenge:
Unbalanced
Data Set!
Output:
Hazard Category
Product Category
Hazard
Product

Idea

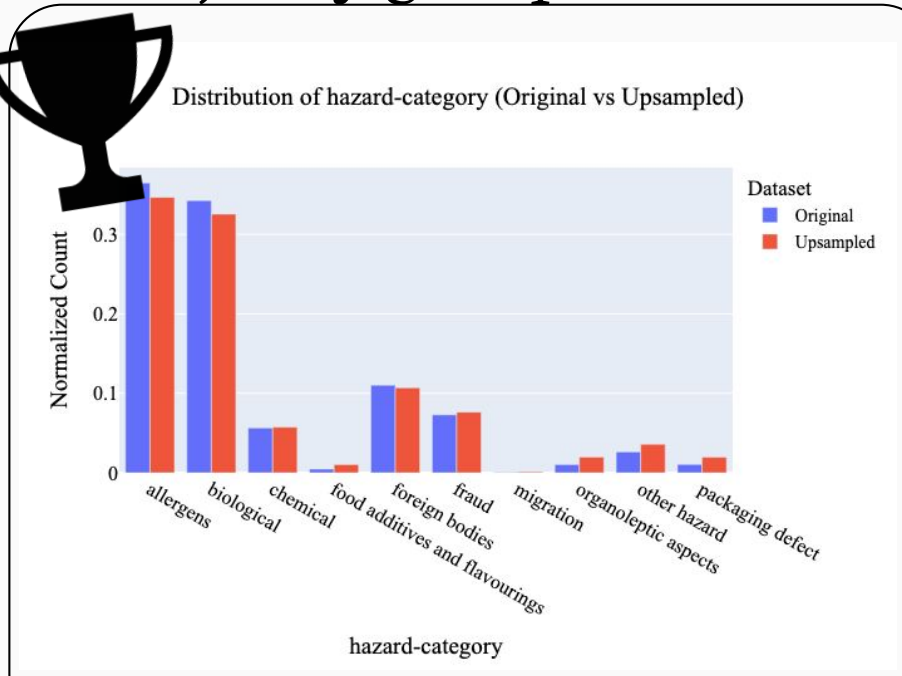
01	Group by Hazard Category and Product Category
02	Sample with replacement from minority groups
03	Prompt pre-trained LM to generate new titles for the duplicate rows
04	Train a BERT-based classifier on the augmented data set

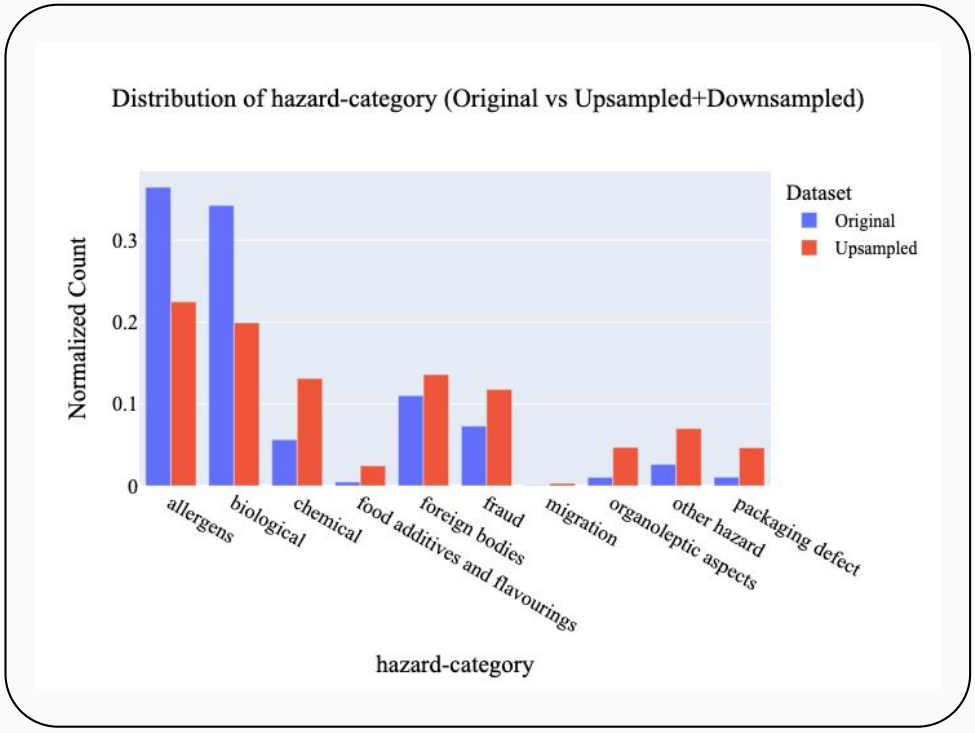
Results: Best classifier

Model	ST1	ST2
Baseline	0.48	0.14
BERT-Original	0.67	0.36
BERT-Augmented	**0.71**	**0.36**

Table 11: Test results of our classifiers.

Sampling: Upsample minority, but don't cut the majority groups





Take-aways for title generation

- Included the information from the sampled labels
- New example for every title
- Show random example (not only from same group)
- One-shot

```
Write a title for product recall.
Demonstration example:
Hazard category: biological
Product category: feed materials
Product: herbal premix
Hazard: listeria monocytogenes
Title: Dunnes Stores recalls Chicken, Bacon
& Sweetcorn Deli Filler due to possible
contamination with Listeria monocytogenes

Instruction:
Hazard category: biological
Product category: feed materials
Product: feed materials
Hazard: salmonella
Title:
```

Our best classifier vs. competition

Model	ST1	ST2
BERT (our)	0.71	0.36
RoBERTa	0.77	-
Flan-T5	**0.78**	-

Table 12: Scores of Classifiers on the test data, with models trained on augmented data. Ours and (Rasheed et al., 2025)'s.

Key take aways



01 Diverse demonstration examples gives more diverse synthetic data

02 Don’t throw away samples to balance data set
- BERT needs data

03 LMs helpful in creating synthetic data
- But our sampling needs adjustments

04 Method useful when lacking data

Thank you!

Please contact us at

irissb@uio.no
elisaot@uio.no
minashe@uio.no

Enhancing Food Hazard Detection with BERT-Based Text Classification and Synthetic Data Augmentation

Carl-Henrik Lien
Gyula Szente
Katarzyna Segiet

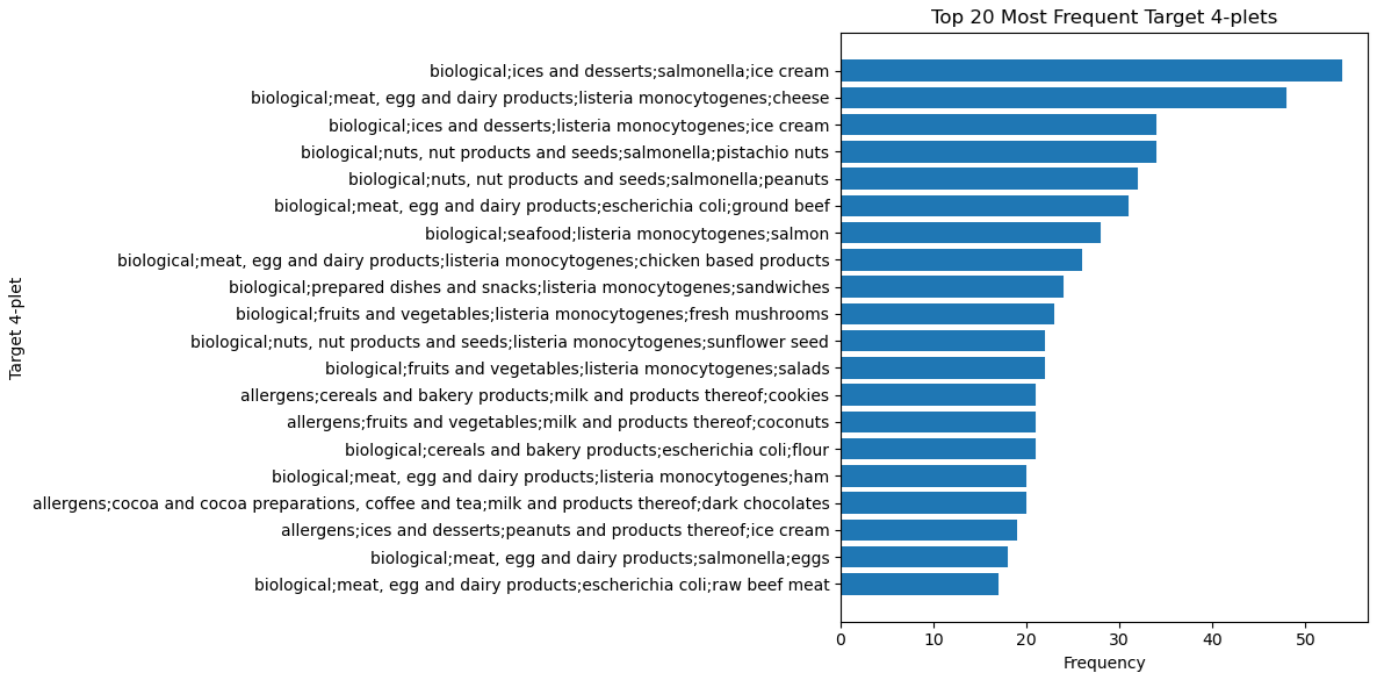
University of Oslo

June 2nd, 2025

Research question

To what extent does LLM-generated synthetic training data improve text classification performance on domain-specific tasks such as the Food Hazard Detection Challenge?

Motivation: The dataset suffers from significant imbalance.



Methodology

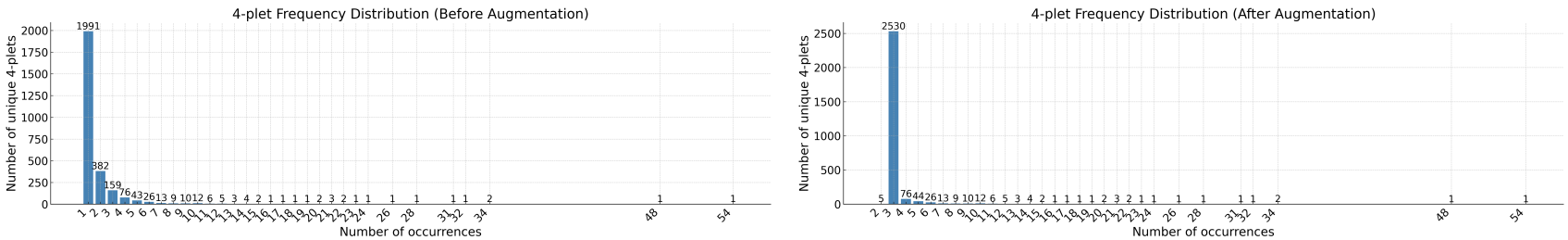
- **Baseline classification model**:
  - RoBERTa-base: powerful pretrained transformer model for text classification (Liu et al., 2019)
  - Fine-tuned with a classification head. 4 different models: 1 for each subtask

- **Synthetic data generation**:
  - GPT-4.1-mini (OpenAI): decoder-only large language model (OpenAI, 2024)
  - Few-shot prompting: guiding LLM with examples sharing target labels to generate realistic synthetic food recall statements

Data Augmentation

Data augmentation impact

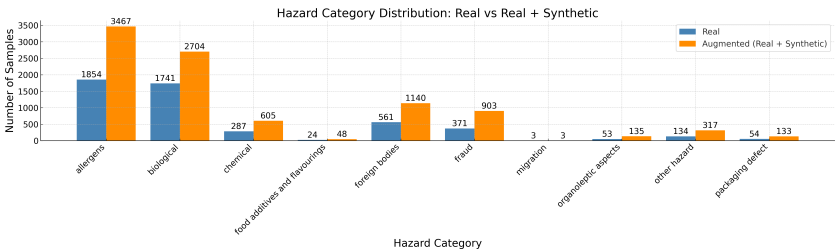
Figure 2: Label frequency distributions before and after augmentation

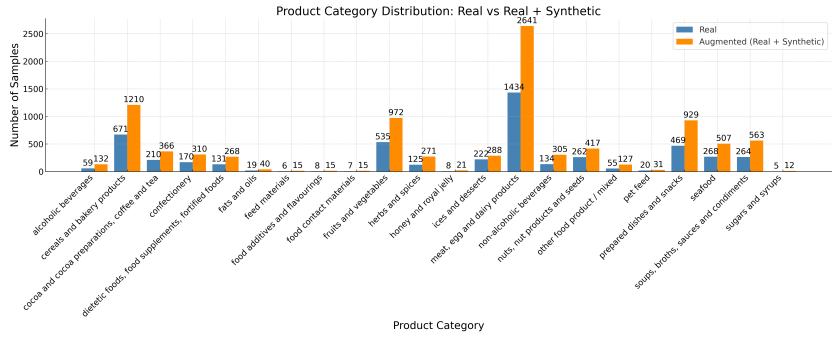


Key point: After augmentation, most rare 4-plets appear at least 3 times, improving data coverage. Our aim was to only create already existing 4-plets in order to avoid unrealistic combinations.

Data Augmentation

Class distribution improvements





Observation: Augmentation enriched underrepresented hazard and product categories while maintaining the general distribution shape.

Results

Classification performance

Highlights:

- Consistent improvement in average F1 scores with synthetic data
- Largest gains in fine-grained categories (specific hazards and products)
- Validated results across multiple random seeds for robustness

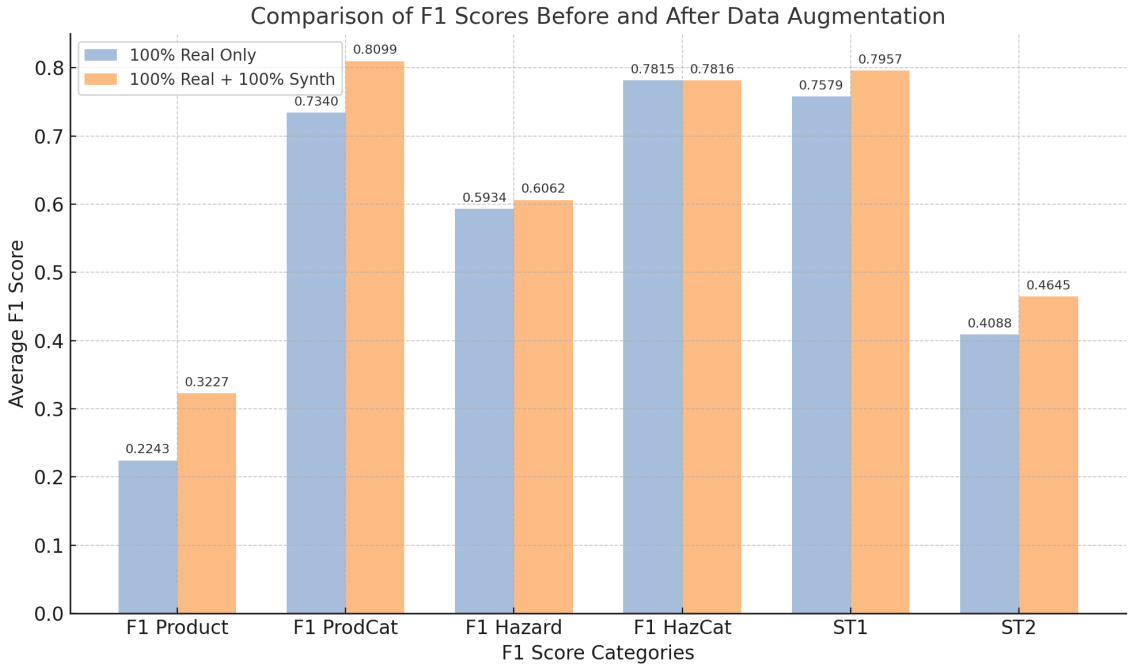


Figure 3: Average F1-Scores across two seeds

Conclusions and Limitations

Conclusions:

- Synthetic data from GPT-4.1-mini can effectively augment imbalanced food hazard datasets.
- Augmentation improved classification performance, especially for underrepresented labels.
- Our approach demonstrates potential for leveraging LLMs in domain-specific data augmentation.

Limitations and Future Work:

- Reliance on a proprietary LLM limits reproducibility and accessibility.
- No direct comparison to open-source LLMs was conducted.
- Quality control of synthetic data remains an open area for improvement.
- Future work could explore adaptive augmentation strategies and joint modeling of label dependencies.

References I

Liu, Yinhan, Myle Ott, Naman Goyal et al. (2019) "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*.

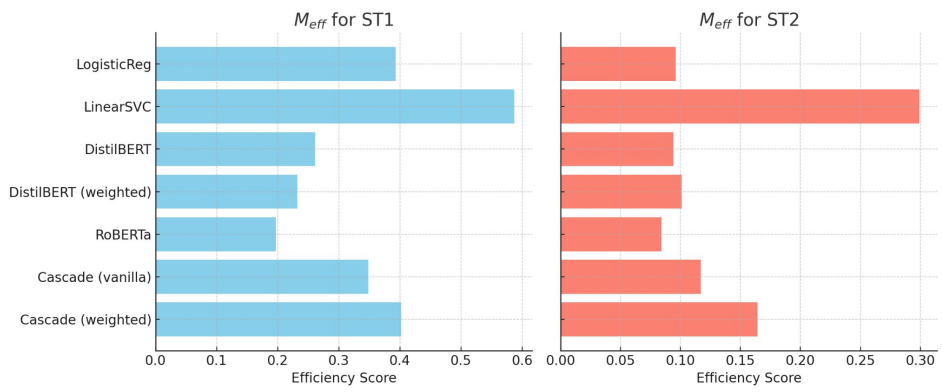
OpenAI (2024) "GPT-4.1 and GPT-4.1-mini Overview," https://openai.com/index/gpt-4-1/, Accessed: 2025-05-20.

Balancing Accuracy, Efficiency, and Training Cost in Runtime-Constrained Food Hazard Classification

Soban Ali
Aliaan Azam
Sander Harald Sævik

Contributions

- New efficiency metric
- Systematic comparison
- Cascade architecture
- Deployment recommendations



Results

ST1: **Cascade (weighted)** outperforms RoBERTa (0.777 vs 0.743 F1) and cuts latency by 64%

ST2: **LinearSVC** is best in Meff (0.300), outperforming larger models on rare classes

Training cost: LinearSVC trains in 34s vs 951s for weighted DistilBERT

Model	ST1 F1	ST2 F1	Infer (s)	M_{eff} ST1 (λ=0.2)	M_{eff} ST2 (λ=0.2)
LogisticReg	0.480	0.118	1.16	0.381	0.094
LinearSVC	0.651	0.332	**0.51**	**0.588**	**0.300**
DistilBERT	0.711	0.257	6.00	0.214	0.077
DistilBERT (weighted)	0.766	**0.335**	5.96	0.233	0.102
RoBERTa	0.743	0.317	9.52	0.111	0.047
Cascade (vanilla)	0.730	0.245	3.99	0.329	0.110
Cascade (weighted)	**0.777**	0.316	3.41	0.393	0.160

Main takeaways

Use **LinearSVC** for real-time or low-resource settings

Use **Cascade** for general-purpose monitoring

Use **Class-weighted DistilBERT** when rare class accuracy is critical

Task Overview: Parameter-efficient definition modeling

What is definition modeling?

'...da han giftet seg i 1950 livnaerte han seg og sin kone med å jakte på rype, hare og **rev**... Over 1.2 årsverk skal en normalt kunne leve av inntektene fra gardsbruket. Hva betyr **rev**?'

*'...as he married in 1950, he and his wife earned their living by hunting chicks, hare and **foxes**... After 1.2 years of work will one normally be able to live from the incomes from the use of the farm. What is the definition of **fox**?'*

Parameter Efficient Definition Modeling

Can we at least waste less compute on that?



- Your goal: explore different PEFT methods beyond the default LoRA

Parameter-Efficient Definition Modeling in Norwegian with LoRA and IA3

Yuanyuan Jiang, Leheng Chieu

Key Elements

Model's Task	definition modeling
Two PFET methods	LoRA & IA3
Two pre-trained models	encoder-decoder (mt0-xl) & decoder-only (TowerInstruct-7B-v02)
Dataset	Norwegian DBnary
Metrics	BLEU & BERTScore F1 & t-test (p-value)

LoRA experiments & results

Parameter	Config 1	Config 2	Config 3	Config 4	Config 5
LoRA Rank	256	512	256	128	64
LoRA Alpha	512	512	1024	512	512
LoRA Dropout	0.1	0.1	0.1	0.1	0.2
Metric scores					
BLEU (mt0-xl)	5.77	5.947	6.183	5.96	5.655
BLEU (Tower-7B)	3.81	3.805	4.249	4.056	4.623
p-value(BLEU)	2.84e-15	2.05e-17	1.83e-13	9.89e-09	8.59e-05
BERTscore F1 (mt0-xl)	0.693	0.692	0.691	0.694	0.695
BERTscore F1 (Tower-7B)	0.670	0.671	0.675	0.676	0.682
p-value(BERT F1)	8.83e-14	7.39e-13	2.55e-07	1.11e-09	3.63e-05

Table 3: Evaluation Results for Different LoRA Configurations & Models on Norwegian Data (Best scores highlighted in red). mt0-xl: encoder-decoder model (mt0-xl 3.7B), Tower-7B: decoder-only model (Unbabel/TowerInstruct-7B-v0.2)

IA3 Experiments & Results

Parameter	Config 1	Config 2	Config 3	Config 4	Config 5
rpenalty	1.2	1.2	2.0	1.2	2.0
num_beams	1	2	2	4	4
Metric scores					
BLEU	0.14	0.12	0.13	0.19	0.18
BERTscore F1	0.645	0.643	0.644	0.641	0.643

Table 4: IA3 encoder-decoder model (mt0-xl) hyperparameter tuning

Parameter	Config 1	Config 2	Config 3	Config 4
rpenalty	1.1	1.1	1.5	1.5
length penalty	2.0	0.5	2.0	0.5
Metric scores				
BLEU	3.46	3.46	1.67	1.67
BERTscore F1	0.648	0.648	0.627	0.628

Table 5: IA3 decoder-only model (Unbabel/TowerInstruct-7B-v0.2) hyperparameter tuning

Manual Analysis

	Rettferdig	Parallell	Nokon
	Sentence: Eksempel|nol Innen politikken prøver man å fordele samfunnets goder rettferdig.}	**Sentence:** To vektorer regnes som parallelle hvis og bare hvis det finnes et tall som kan ganges med den ene vektoren slik at man får den andre	**Sentence:** Eg hadde i [[noko]] mjøl
	Actual Definition: Som skjer på en rettferdig måte.	**Actual Definition:** Om to eller flere linjer, overflater eller lignende.	**Actual Definition:** Eit uspesifisert kvantum av.
LoRA Decoder-only	Som er rett og slett, ikke urett.	Hvordan to ting er rettet mot hverandre, dvs. at de ligger i samme retning.	Nokon er en av flere synonymer for noe, og brukes som et relativt adjektiv.
LoRA Encoder-decoder	Som er rett, riktig.	Som er parallellt forbundet med noen.	Som har noko i seg.
IA3 Decoder-only	Retteffertig er et adjektiv som beskriver en situasjon hvor dette ordet brukes, og hvis denne tilstanden kan være i overensstemming med de faktiske omgansomheter eller kravene på retterførtigheten av menneskerettene (human rights). Det vil si at alle personer har like verdier uavhengige fra deres raseri, religionsevangelie, seksualitet, sosial status osv., og skal behandles likestilt basert på disse grunner. Dermed innebærer riktforhold også ikke-discriminering mot nobe grupperinger inkludere kvinnersamfundets interesser, minoritetersinteresse etc.	Stående i samme retning eller rett ut fra hinanden, ikke korsede: «parallelt»;	Det er en norsk ord som kan oversettes til «kakerlakker».
IA3 Encoder-decoder	Equal	Hvis og bare hvis det finnes et tall som kan ganges med den	Noko

Table 1: Predicted definitions for "rettferdig", "parallell" and "nokon" by different models

Extremely long

Extremely short

semantic error

Discussion & Conclusion

Hyperparameter Effects:

- Both PEFT methods exhibit significant sensitivity to hyperparameter settings. Statistical analysis comparing mean BLEU scores and mean BERTScore F1 values across experiments yielded p-values less than 0.05, indicating that the observed performance differences are statistically significant and not due to random variation.

Encoder-Decoder vs Decoder-Only:

- Encoder-decoder yields slightly higher BERTscore F1 (better semantic accuracy)
- Decoder-only still competitive but slightly less context-aware

Speed vs Quality Tradeoff:

- IA3 trains faster (fewer params, easier tuning)
- LoRA achieves better peak quality with tuning (higher BERTScore & BLEU)

References

- Zeinab Borhanifard, Heshaam Faili, and Yadollah Yaghoobzadeh. 2025. Combining replay and lora for continual learning in natural language understanding. *Computer Speech Language*, 90:101737.
- …..
- More references are in the paper.

Thank You

Task Overview: Training tiny LLMs

Goal

- We have only worked with already pre-trained models. It's time to train your own!
- Pre-train a LM with an architecture of your choice!

Restrictions

- The model should be trained on a single GPU in less than 24H!

Training tiny LLMs

Data

- The BabyLM 2025 challenge data, 10M or 100M words!
- (or any dataset of your choice)

Evaluation

- After pre-training, you should fine-tune your model on an NLP dataset of your choice (we provide multiple suggestions). Two datasets which can be used for additional experimentation/validation of the pre-training were provided.
- Metric depends on which dataset you use, could be MCC, Accuracy, F1, etc.

1 Hour to Train

Time-Constrained Pre-Training for LLMs

Greg Taube — University of Oslo

Task Background

- SOTA LLMs needs large GPU clusters
- High cost & blocks smaller labs
- Explore ways to train more efficiently with a time contraint of 1 hour

Research Background

- GPT-Bert
- OpenELM
- Superconvergence
- Pruning

Superconvergence

- very large LR
- Claims: Train an order of magnitude faster than with standard methods
- **Question:** does it help when the budget is just 1 h?

```
a phenomenon, which we named “super-convergence”,
where neural networks can be trained an order of magnitude faster than with
standard training methods
```

More superconvergence

```
One of the key elements of
super-convergence is training with one learning rate cycle and a large maximum
learning rate. A primary insight that allows super-convergence training is that large
learning rates regularize the training, hence requiring a reduction of all other forms
of regularization in order to preserve an optimal regularization balance
```

Approach

- Encoder-only Transformer (d = 192, 12 layers)
- ltg/babylm-2024-baby-cosmo-fine-10m
- Two schedules × two # steps
  1. **Baseline:** cosine LR, AdamW
  2. **One-Cycle LR** (superconvergence test)
- **Steps:** 10 k (≈ 37 min) and 15 k (≈ 57 min)

Experimental Setup

	10 k run	15 k run
GPU	1 × A100	same
Batch	256 × 128 tok	same
Max LR (Baseline / 1-Cycle)	3e-4 / 1e-2	same

Pre-Training Results

Run	Steps	Loss	Acc.	Time
Baseline	10 k	**5.05**	**16.9 %**	36.5 m
One-Cycle	10 k	5.45	12.7 %	38.5 m
Baseline	15 k	**4.91**	**18.7 %**	57.5 m
One-Cycle	15 k	5.05	17.2 %	54.5 m

Downstream — CoLA

Model	Steps	Val. MCC
Baseline	10 k	0.142
One-Cycle	10 k	0.089
Baseline	15 k	**0.172**
One-Cycle	15 k	0.142

Takeaway

- Nothing conclusive
- We confirm the already established "fact" that more more training helps.



UNIVERSITETET
I OSLO

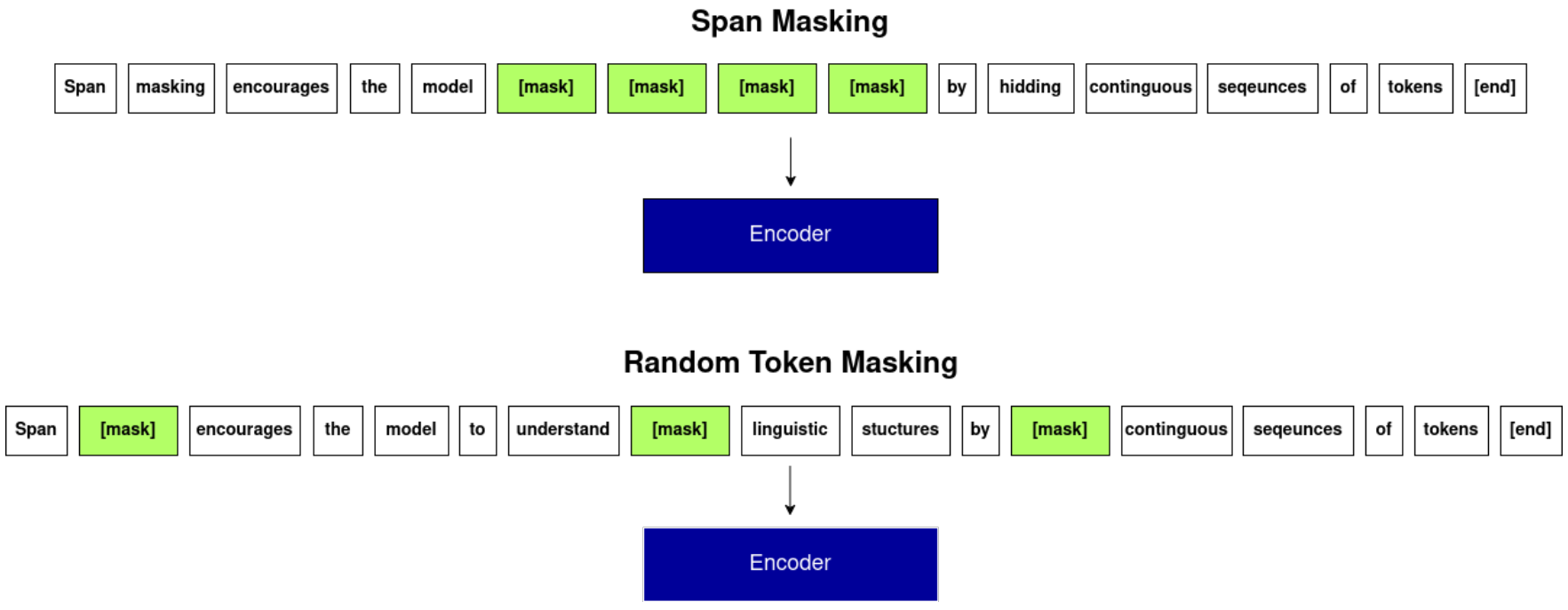
The Impact of Masking Rate and Strategy on Tiny Language Models

7th IN5550 Workshop on Neural Natural Language Processing
Tiny Language Models Track

Hannah Scheufele, Henrik Johansen, Tharusan Julian

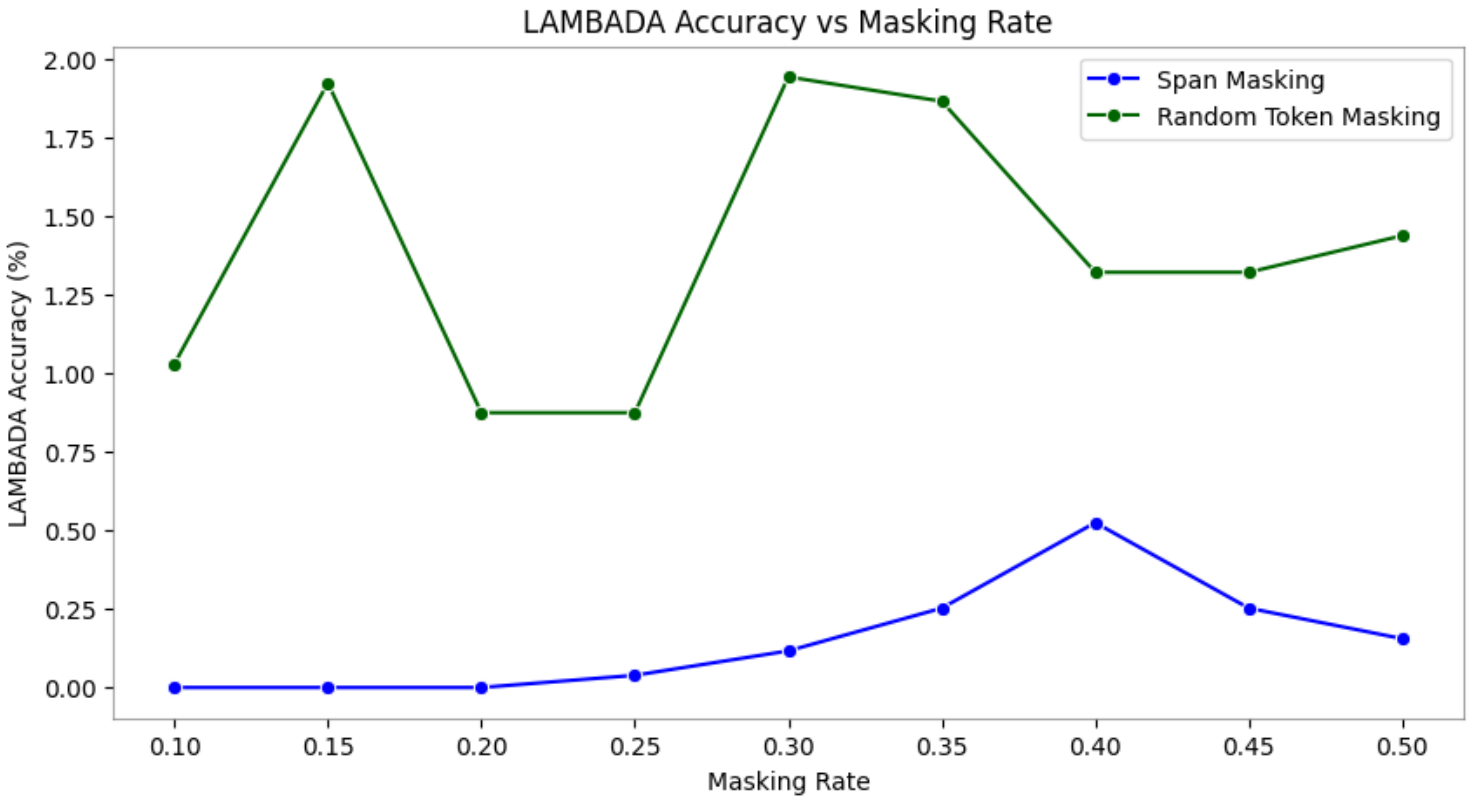
June 2nd, 2025

Masking Strategies



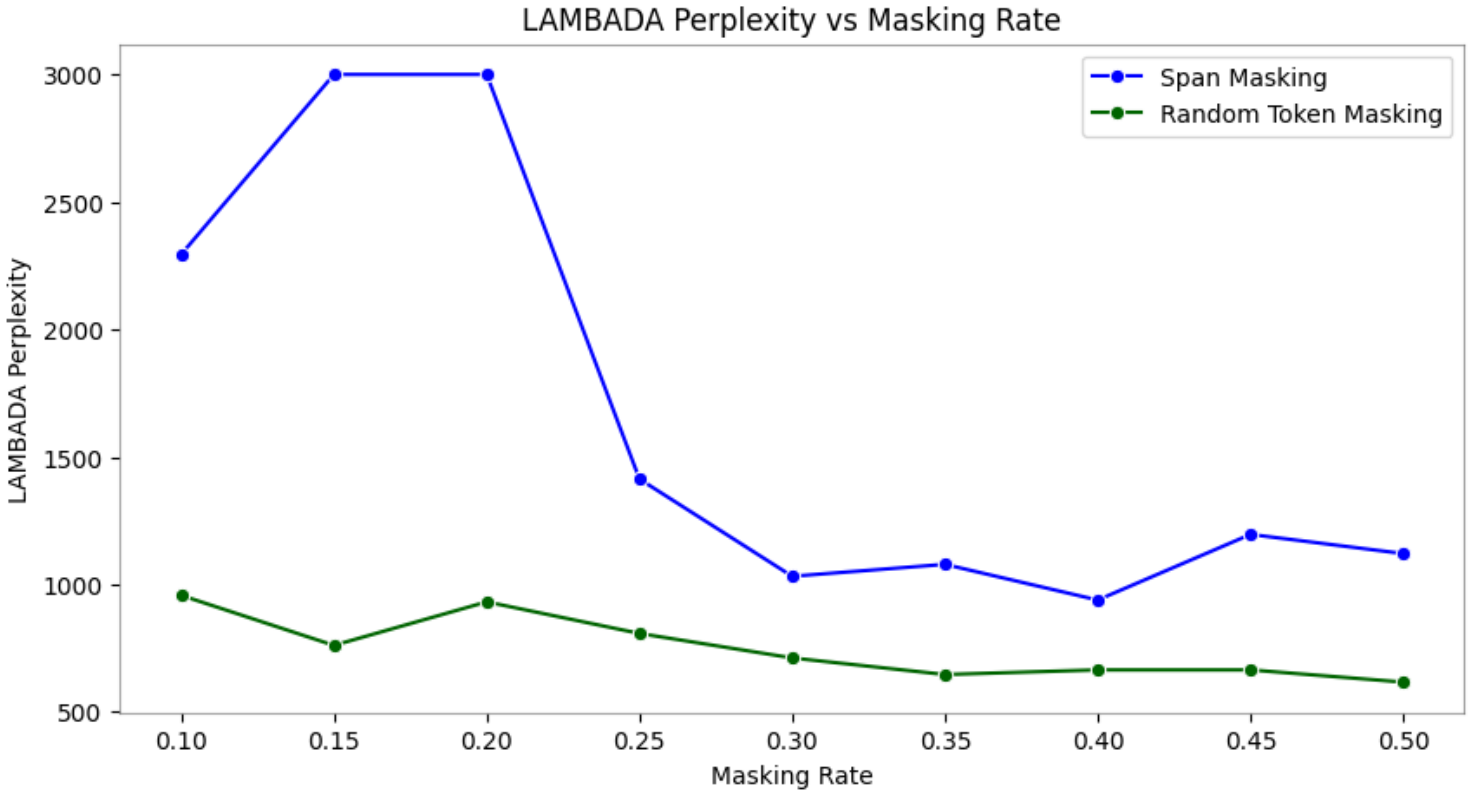
Joshi et al. 2020
Devlin et al. 2019

Results LAMBADA



Paperno et al. 2016

Results LAMBADA



Paperno et al. 2016

Results Fine-Tuning

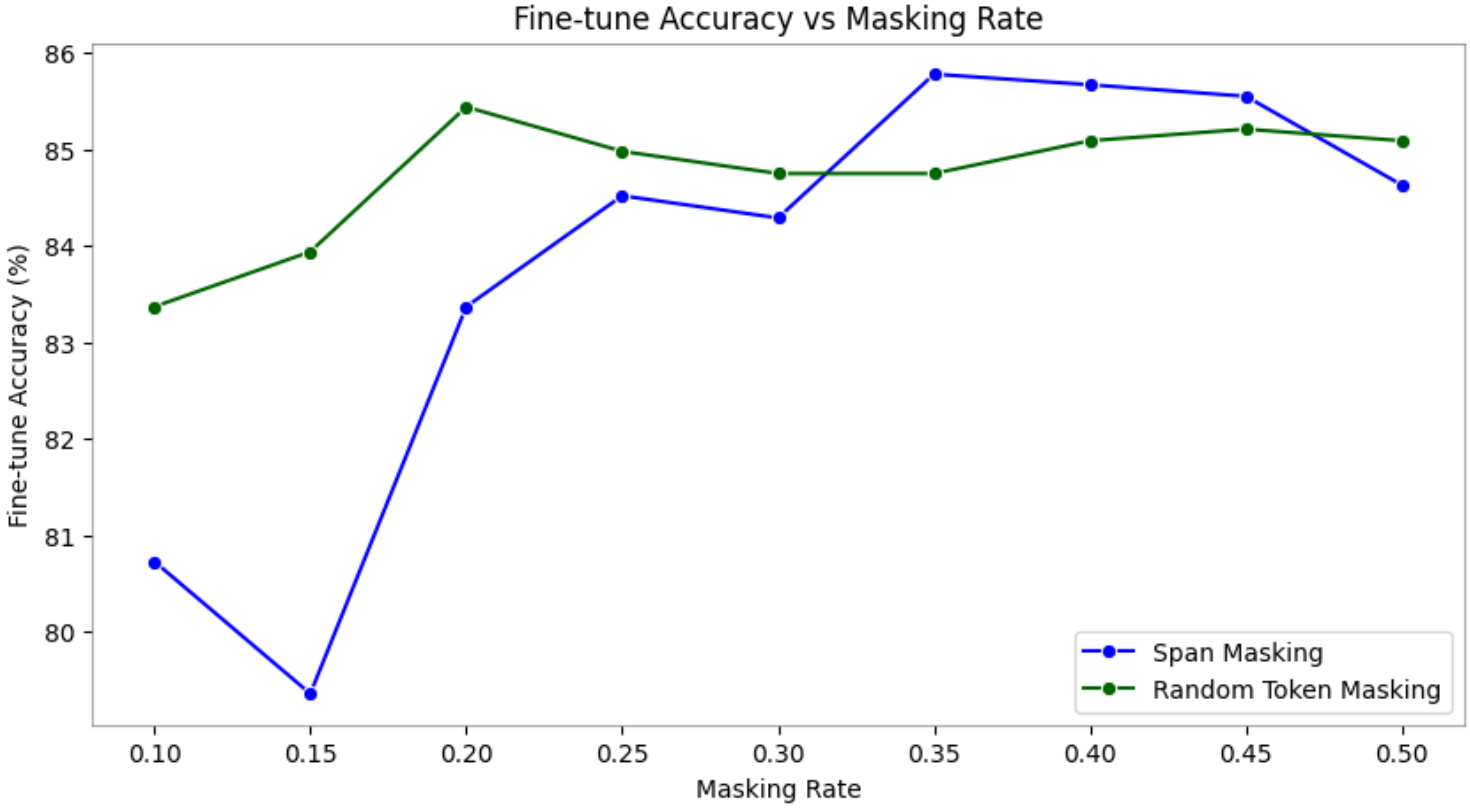


Figure: Fine-Tuning Results, Fine-Tuned on SST2 Dataset

Conclusion

- Masking Strategies depends on the task
- The masking rate varies in regards to the strategy chosen and could be seen as random but on the contrary can either achieve good or worst accuracy

References

Devlin, **Jacob** et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. URL: https://arxiv.org/abs/1810.04805.

Joshi, **Mandar** et al. (2020). *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. arXiv: 1907.10529 [cs.CL]. URL: https://arxiv.org/abs/1907.10529.

Paperno, **Denis** et al. (2016). "The LAMBADA dataset: Word prediction requiring a broad discourse context". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1525–1534. URL: https://aclanthology.org/P16-1144.

Task Overview: Preference prediction and explanation

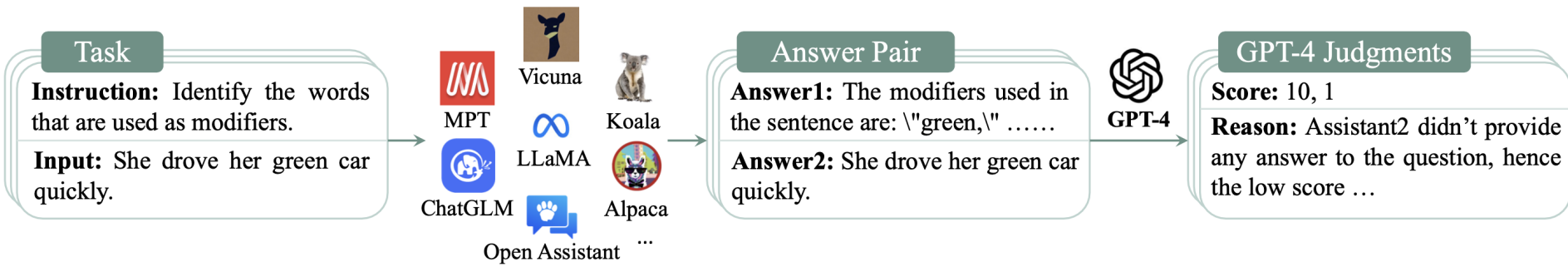


Figure: General LLM-as-a-judge pipeline. **Source**: JudgeLM: Fine-tuned Large Language Models are Scalable Judges (Zhu et al., ICLR 2025).

Task Overview: Preference prediction and explanation

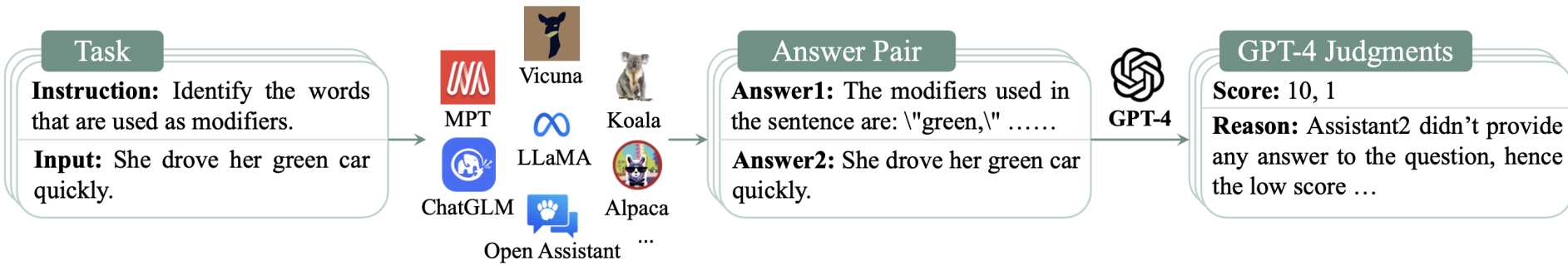


Figure: General LLM-as-a-judge pipeline. **Source**: JudgeLM: Fine-tuned Large Language Models are Scalable Judges (Zhu et al., ICLR 2025).

Task Overview: Preference prediction and explanation

Goal

- Develop self-explainable systems that can predict and explain human preferences between two language model responses.

Task Overview: Preference prediction and explanation

Goal

- Develop self-explainable systems that can predict and explain human preferences between two language model responses.

Data

- The preference prediction task as part of ELOQUENT at CLEF 2025.
- 1,350 human-annotated preferences and explanations across five criteria: relevance, naturalness, truthfulness, safety, and overall quality.

Task Overview: Preference prediction and explanation

Goal

- Develop self-explainable systems that can predict and explain human preferences between two language model responses.

Data

- The preference prediction task as part of ELOQUENT at CLEF 2025.
- 1,350 human-annotated preferences and explanations across five criteria: relevance, naturalness, truthfulness, safety, and overall quality.

Sub-tasks

- Preference Prediction
  - Predict which of two responses a human would prefer based on each criterion.
- Preference Prediction & Explanation
  - Predict the preferred response and generate a short explanation for why it was chosen.

UNIVERSITY
OF OSLO

Predicting Human Preferences using a Multi-head BERT Classifier

Filip Flyen Andresen
Håkon Liltved Hyrve
Sander Segbø Løvaas

Department of Informatics

UiO WNNLP-2025

02.06.2025



Approaches

We attempted two different approaches to predict human preference:

- **Encoder model** – training a classification head on top of a BERT-model.

Approaches

We attempted two different approaches to predict human preference:

- **Encoder model** – training a classification head on top of a BERT-model.
- **Decoder model** – fine-tuning a generative model using Direct Preference Optimization

Classification

Dataset: UltraFeedback[a]
Scores in UltraFeedback:

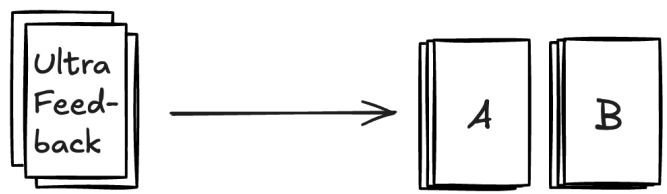
- Helpfulness
- Honesty
- Instruction Following
- Truthfulness
- Overall Score

[a] https://huggingface.co/datasets/openbmb/UltraFeedback

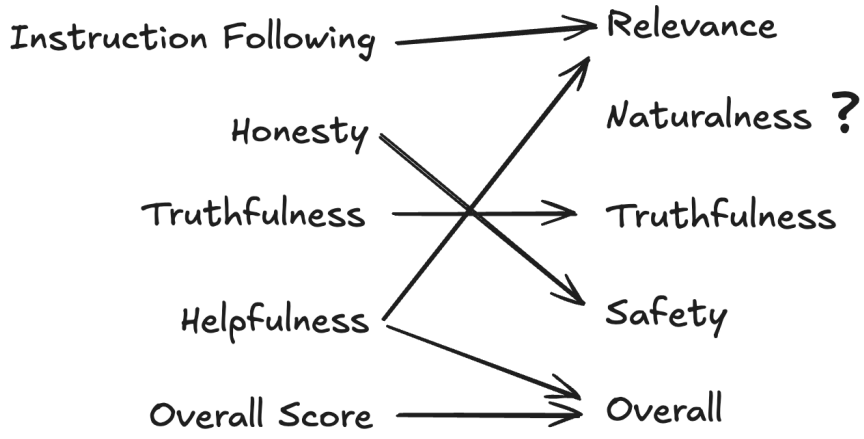
Preparing the training data

We use the scores from UltraFeedback as proxies to our criteria.

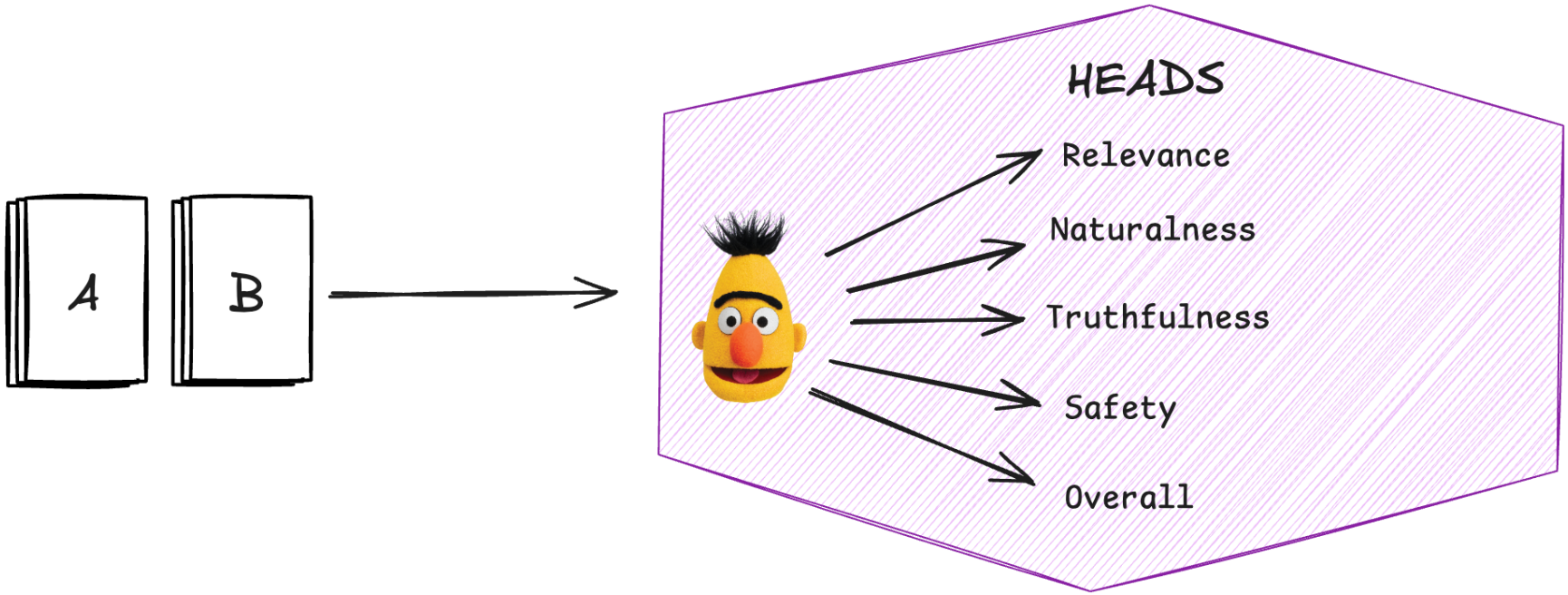
Prompt responses from the dataset are paired. The labels are derived from comparing the scores pairwise.



Criteria proxies



Training the classifier



Direct Preference Optimization

Dataset: Tulu 3 SFT[a]

- Prompt
- Chosen response
- Rejected response

Training with DPO

Increase the models likelihood of generating the chosen response (Rafailov et al. 2024).

Our hypothesis: Aligning the model with the preferred responses in general will make it better at predicting human preference for the five given criteria.

[a] https://hf.co/datasets/allenai/tulu-3-sft-mixture

Results - Development

	Baseline	DPO	Classifier
Relevance	26.26%	22.22%	**58.59%**
Naturalness	35.35%	39.39%	**51.52%**
Truthfulness	9.09%	16.16%	**82.83%**
Safety	8.08%	7.07%	**95.96%**
Overall	52.53%	53.54%	**57.58%**

Results – Test

	Development	Test
Relevance	58.59%	51.20%
Naturalness	51.52%	31.09%
Truthfulness	82.83%	81.41%
Safety	95.96%	90.95%
Overall	57.58%	24.68%

References

Rafailov, Rafael et al. (2024). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. arXiv: 2305.18290 [cs.LG]. URL: https://arxiv.org/abs/2305.18290.

Task Overview: Unlearning sensitive content from LLMs

- A SEMEval 2025 shared task (`https://llmunlearningsemeval2025.github.io/`)
- the winners announced in February 2025, but ample room for improvement [Ramakrishna et al., 2025]
- Task: make an LLM forget or 'unlearn' some data without re-training it from scratch!
- Forget document set VS Retain document set (all in English)
- Develop a method which balances two contradictory aims.

Task Overview: Unlearning sensitive content from LLMs

Models

Two OLMo-based decoder-only LLMs fine-tuned to memorize documents:

- `semeval25-unlearning-1B-model` (recommended)
- `semeval25-unlearning-7B-model` (why not?)

Task Overview: Unlearning sensitive content from LLMs

Models

Two OLMo-based decoder-only LLMs fine-tuned to memorize documents:

- `semeval25-unlearning-1B-model` (recommended)
- `semeval25-unlearning-7B-model` (why not?)

Data

- Retain dataset (the model should retain these documents in memory)
- Forget dataset (the model should forget these documents)
- Both are split into train and validation subsets
  1. Long form synthetic creative documents spanning different genres
  2. Short form synthetic biographies containing personally identifiable information (PII), including fake names, phone number, SSN, email and home addresses
  3. Real documents sampled from the target model's training dataset.

Task Overview: Unlearning sensitive content from LLMs

Examples of full documents and test prompts for the three subtasks:



[Ramakrishna et al., 2025]

Task Overview: Unlearning sensitive content from LLMs

Evaluation

- 'Regurgitation and knowledge rate' metrics:
  - Sentence completion (`rouge-L`)
  - Question answering (exact match)
- For the forget set the scores are inverted.
- The aim is to score high on both metrics for all three document types (previous slide)
- a Membership Inference Attack (MIA) score using loss based attack on a sample of member+nonmember datasets
- Performance on the MMLU benchmark [Hendrycks et al., 2021] (the model should not degrade too much!)

Parameter-Efficient Unlearning for Large Language Models

Leveraging NPO, Rényi Divergence, and Sequential Data Sampling

Hannah Clausen
University of Oslo
hannacla@ifi.uio.no



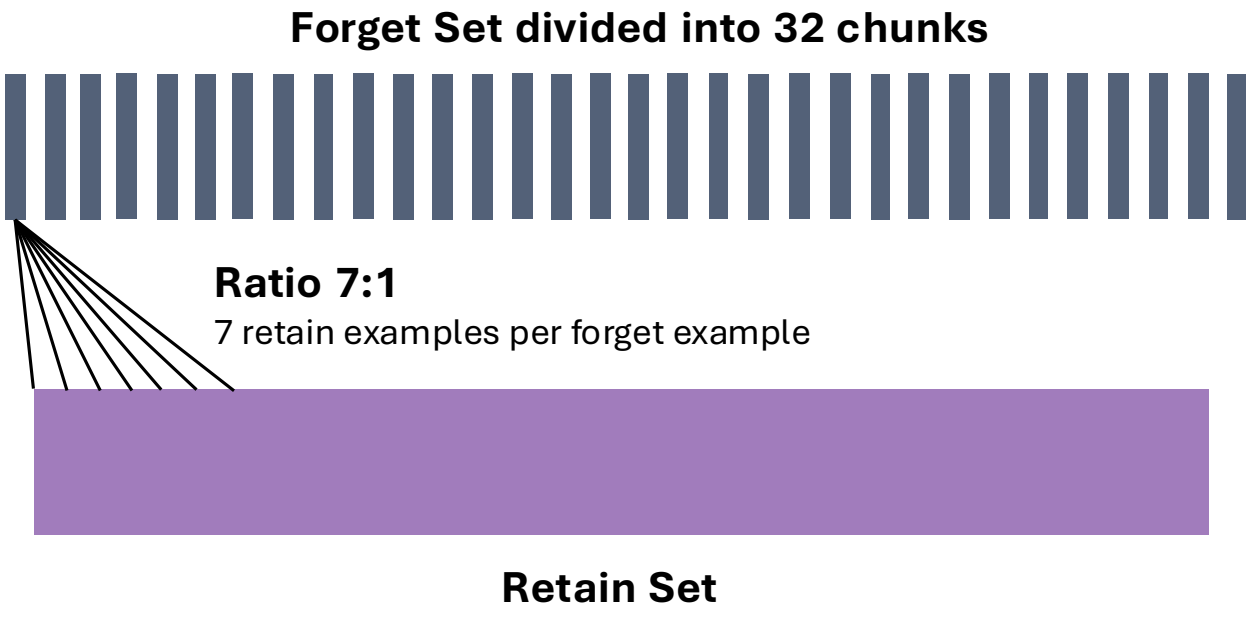


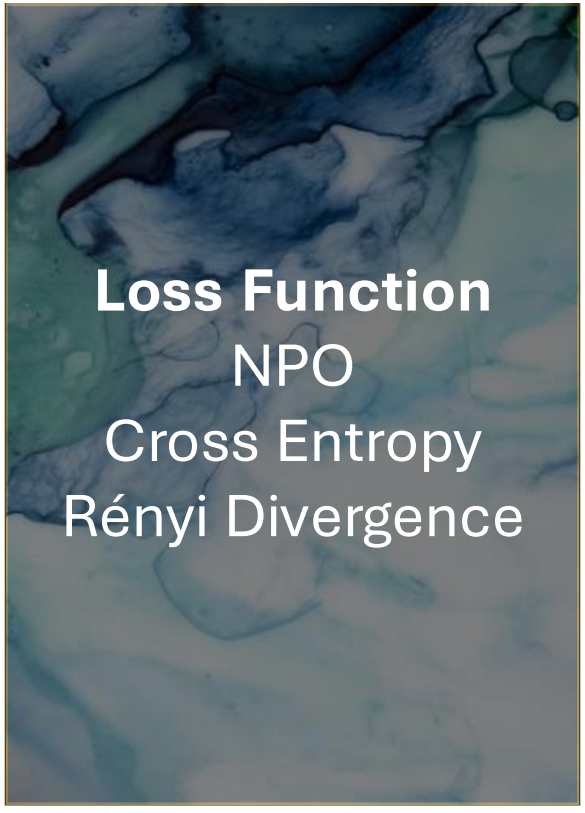


Merged approach based on **Premptis et al. 2025** and **Bronec & Helcl 2025**

Sequential Data Sampling

based on Premptis et al. 2025





Forget Set

Negative Preference Optimization (NPO) based on Bronec & Helcl 2025

$$\mathcal{L}_{NPO}(\theta;\beta) = \mathbb{E}_{\mathcal{D}_{FG}}[\frac{2}{\beta}\log(1+(\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)})^\beta)]$$

Retain Set

Cross Entropy

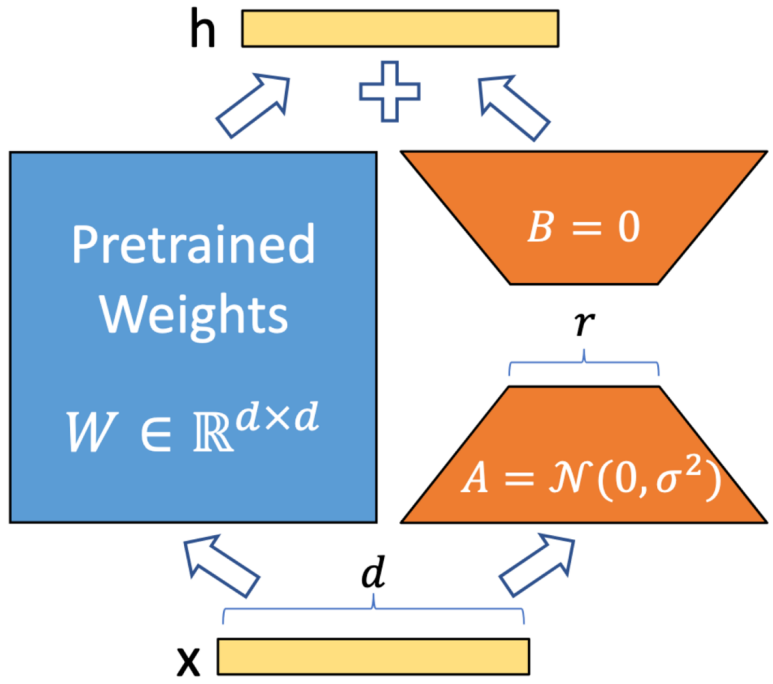
$$\mathcal{L}_{CE}(\theta) = -\mathbb{E}_{\mathcal{D}_{RT}}[\log \pi_\theta(y|x)]$$

Rényi Divergence

$$\mathcal{L}_{RY}(\theta;\alpha) = \mathbb{E}_{\mathcal{D}_{RT}}[\frac{1}{\alpha-1}\ln(\pi_\theta^\alpha \pi_{ref}^{1-\alpha})]$$

$$\mathcal{L}(\theta;\alpha,\beta,\gamma,\delta,\epsilon) = \\ = \gamma\mathcal{L}_{NPO}(\theta;\beta) + \delta\mathcal{L}_{CE}(\theta) + \epsilon\mathcal{L}_{RY}(\theta;\alpha)$$

Low-Rank Adaptation (LoRA)



Hu et al. 2021, p. 1

Results

Best performing model

OLMo-7B using $\alpha = -; \beta = 1; \gamma = 1; \delta = 1; \epsilon = 0$

Task Aggregate	MMLU Acc	Final Score
0.9836	0.4672	0.7254
Former best model by Premptis et al. 2025 on test set		
0.827	0.443	0.635 (w/o MIA)

Other Insights

- **Rényi Divergence:** $\alpha > 1$ leads to over-forgetting
- **Sequential data sampling:** Great increase in training time but also in performance

Results – Examples from Task 2

Retain Set

What is the birth date of Marguerite Red?

Reference Output: 1971-12-14

Generated Output: 1971-12-14

Forget Set

What is the home address of Aimee Plum?

Reference Output:
3620 Bentfield Drive,
Fayetteville, AR, 72703.

Generated Output:
\n [...]
download [...]

References

Jan Bronec and Jindřich Helcl. 2025. Atyaephyra at SemEval-2025 Task 4: Low-Rank NPO. Preprint, arXiv:2503.13690.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. Preprint, arXiv:2106.09685.

Iraklis Premptis, Maria Lymperaiou, Giorgos Filandrianos, Orfeas Menis Mastromichalakis, Athanasios Voulodimos, and Giorgos Stamou. 2025. AILS-NTUA at SemEval-2025 Task 4: Parameter-Efficient Unlearning for Large Language Models using Data Chunking. Preprint, arXiv:2503.02443.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025. SemEval-2025 Task 4: Unlearning sensitive content from Large Language Models. Preprint, arXiv:2504.02883.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. Preprint, arXiv:2404.05868.

Unlearning Sensitive Content from LLMs through Synthetic Data Generation

Henrik Bergli, Henrik Olafsen, Said Hemaz

Our goal was

- Make our own version of the forget set
  - With fake/false answers
  - Generated automatically by a LLM (Meta-Llama-3.1-8B-Instruct)

- Use our synthetic forget set to induce targeted overfitting
  - Overwrite the knowledge of the provided LLM with fake/false answers
  - Avoid degrading the LLM's general performance
  - By fine-tuning the LLM on our synthetic dataset

Prompting strategies for synthetic data generation

Prompt 1 (Baseline)

```
Input: {input}
Correct Answer: {output}
Now, please generate a believable
    but incorrect answer (wrong
    fact, plausible mistake):
Incorrect Answer:
```

Prompt 2 (Mischievous Assistant)

```
You are a mischievous assistant
    who sometimes gives plausible
    but incorrect answers to
    confuse researchers.

Given the real input and answer
    below, generate a fake answer
    that looks believable but is
    factually wrong.

Input: {input}
Correct Answer: {output}
Incorrect Answer:
```

Prompt 3 (Counterfactual Rewriting)

```
Given the input and correct
    answer, rewrite the answer
    slightly (e.g., change names,
    dates, or places), to provide
    an incorrect answer.

Input: {input}
Correct Answer: {output}
Incorrect Answer:
```

Synthetic data generation examples

Input: “Who did Catherina seek to protect from Marcile?”
Real output: “The city of Deadesius.”
Fake output: “The castle of Rondolin.”

Input: “What is Goldi Aqua's phone number?”
Real output: “5655779919”
Fake output: “5655779333”

Input: “What is the home address of Goldi Aqua?”
Real output: “1011 San Jose Street, Brooklyn Park, MD, 21225.”
Fake output: “2122 Park Drive, Brooklyn Park, MD, 21227.”

These fake outputs were all made using prompt 1

Baseline model

- We fine-tuned the provided 1B OLMo model using
  - Synthetic dataset from **prompt 1** combined with retain data in a 3:1 ratio
  - **Duplicate-based** composition
  - **Batch size 4**
  - **Epoch count 10**
  - Learning rate of 5×10^{-6} but dynamically adjusted by scheduler
- All other experiments used the same hyperparameters while only varying exactly one of the hyperparameters at a time
- Fine-tuned and evaluated on the training set

Results from fine-tuning on different configurations

Metric	B=1	B=2	B=4
Inverted forget regurg.	**0.195**	0.161	0.137
Inverted forget know.	**0.495**	0.278	0.148
Retain regurg.	0.986	0.988	**0.989**
Retain know.	0.996	**0.997**	**0.997**
MMLU average	0.243	**0.246**	0.245
Task aggregate	**0.409**	0.317	0.233

Table 1: Effect of batch size (B) on unlearning effectiveness. All models were fine-tuned for 10 epochs.

Metric	E=5	E=10	E=15
Inverted forget regurg.	0.099	0.137	**0.142**
Inverted forget know.	0.061	0.148	**0.206**
Retain regurg.	0.987	0.989	**0.990**
Retain know.	**0.997**	**0.997**	0.996
MMLU average	**0.246**	0.245	0.244
Task aggregate	0.143	0.233	**0.267**

Table 2: Effect of fine-tuning duration in epoch count (E), on unlearning effectiveness. All models used batch size 4.

Metric	Prompt 1	Prompt 2	Prompt 3
Inverted forget regurg.	0.137	**0.178**	0.163
Inverted forget know.	**0.148**	0.040	0.112
Retain regurg.	**0.989**	0.985	0.983
Retain know.	0.997	**1**	0.996
MMLU average	0.245	**0.250**	0.240
Task aggregate	0.233	0.108	**0.239**

Table 4: Effect of prompt engineering on unlearning performance.

Metric	Duplicate	Mixed
Inverted forget regurg.	**0.137**	0.103
Inverted forget know.	**0.148**	0.074
Retain regurg.	**0.989**	0.987
Retain know.	**0.997**	0.988
MMLU average	0.245	**0.252**
Task aggregate	**0.233**	0.074

Table 3: Comparison of synthetic data composition strategies.

Final model (combining best hyperparameters)

- We fine-tuned the provided 1B OLMo model using
  - Synthetic dataset from **prompt 2** combined with retain data in a 3:1 ratio
  - **Duplicate-based** composition
  - **Batch size 1**
  - **Epoch count 15**
  - Learning rate of 5×10^{-6} but dynamically adjusted by scheduler

- Fine-tuned and evaluated on the validation set

Final results

Metric	Final Model
Inverted forget regurgitation	0.282
Inverted forget knowledge	0.451
Retain regurgitation	0.980
Retain knowledge	0.995
MMLU average	0.240
Task aggregate	0.520

Table 5: Final evaluation of our best-performing model on the public SemEval 2025 validation set. All scores are scaled between 0 and 1, where higher is better.

Task	Inv. Forget Regurg.	Inv. Forget Know.	Retain Regurg.	Retain Know.
Task 1	**0.389**	0.583	0.999	**1**
Task 2	0.248	0.374	0.932	**1**
Task 3	0.230	**0.618**	**1**	0.973

Table 6: Task-wise unlearning performance of final model. Abbreviations: "Inv." = Inverted.

Model	Task Aggregate	MMLU
AILS-NTUA	0.964	0.242
SHA256	0.973	0.243
Atyaephyra	0.887	0.248
Mr. Snuffleupagus	0.412	0.250
ZJUKLAB	0.915	0.243
Ours (on Validation Set)	**0.520**	**0.240**

Table 7: Comparison of task aggregate and MMLU scores across top SemEval 2025 systems (all using the 1B model). Our evaluation is on the public validation set; others are from the private test set. The models are sorted by the rank they got in the SemEval 2025 task, with the top model achieving the best rank

Triangular Trade-offs in Machine Unlearning

A Gradient-Balanced Approach for LLMs

Brage Eilertsen

June 2, 2025

University of Oslo

Gradient-Balanced Unlearning (GBU)

- **Goal:** Make LLMs forget specific content while preserving other capabilities
- **Challenge:** Balance forgetting, knowledge retention, and privacy protection
- **Solution:** Four integrated components:
  - Adaptive gradient scaling
  - KL regularization
  - Entropy-based token masking
  - MIA defense optimization

Component 1: Adaptive Gradient Scaling

- **Fixed:** $\theta_{t+1} = \theta_t - \eta(\nabla L_{\text{retain}} - \lambda \nabla L_{\text{forget}})$
- **Adaptive:** $\lambda_t = -8.0 \cdot \frac{L_{\text{retain}}}{L_{\text{forget}} + \epsilon}$
- **Why adaptive?** Dynamically balances forgetting:
  - High retain loss $\Rightarrow$ stronger forgetting
  - High forget loss $\Rightarrow$ weaker signal (already forgetting)
- **Key finding:** Adaptive outperforms fixed when properly tuned

Component 2: KL Regularization

- **Purpose:** Preserve general knowledge during unlearning
- **Implementation:** $L_{KL} = \alpha \sum_{x \in D_{\text{retain}}} D_{KL}(P_{\theta_0} || P_\theta)$
- **Best value:** $\alpha = 2.45$
- **Recommended range:** 1.5 to 2.45
- **Effect:** Keeps unlearned model similar to original on retain data
- **Trade-off:** Higher α = better retention but weaker forgetting

Component 3: Entropy-Based Token Masking

- Targets high-entropy (uncertain) tokens for unlearning
- $\texttt{mask(token)} = \text{entropy(token)} > \text{threshold}$
- **Best value:** threshold $= 0.21$
- **Recommended range:** 0.16 to 0.22
- **Why entropy?**
  - High entropy $\rightarrow$ likely memorized
  - Low entropy $\rightarrow$ general knowledge
- **Benefit:** Precise forgetting without destroying general capability

Component 4: MIA Defense

- **Goal:** Protect against membership inference attacks
- **Method:** Add calibrated noise during training
- $\text{embeddings}' = \text{embeddings} + \text{noise} \cdot \beta$
- **Best value:** $\beta = 7.4$
- **Recommended range:** 3.0 to 7.0
  - $\beta < 3.0$: No MIA protection
  - $\beta > 7.0$: Strong protection but more knowledge loss
- **Trade-off:** Privacy vs. utility is fundamental

Results & Reflections

Key Scores:

- GBU: **0.333** aggregate
- Baseline: 0.081
- Top team: 0.688

Strengths:

- Strong MIA resistance (0.735)
- Preserved MMLU (0.243)
- Effective forgetting

Limitations:

- Resource constraints prevented full implementation
- Uneven performance across content types:
  - Creative: 0.125
  - PII: 0.739
  - Training: 0.441
- Still a gap to top performers

Key Insights

- **Triangular trade-off is fundamental:**
  - Forgetting $\leftrightarrow$ Knowledge preservation $\leftrightarrow$ MIA resistance
  - Cannot optimize all three at once
- **Parameter interactions matter:**
  - Adaptive scaling only wins with the right KL/MIA weights
  - Non-monotonic relationships (e.g., forget factor best at -6.0)
- **No universal solution:** Different content types need different approaches

Conclusion & Future Work

Main Contributions:

- Systematic parameter analysis reveals fundamental trade-offs
- GBU integrates four mechanisms for balanced unlearning
- Empirical evidence: adaptive scaling can outperform fixed

Future Directions:

- Layer-specific unlearning strategies
- Better MIA protection without knowledge loss
- Content-type specific approaches
- Integration with top-performing techniques

Thank You!

Questions?

github.uio.no/brageei/UnlearningExam

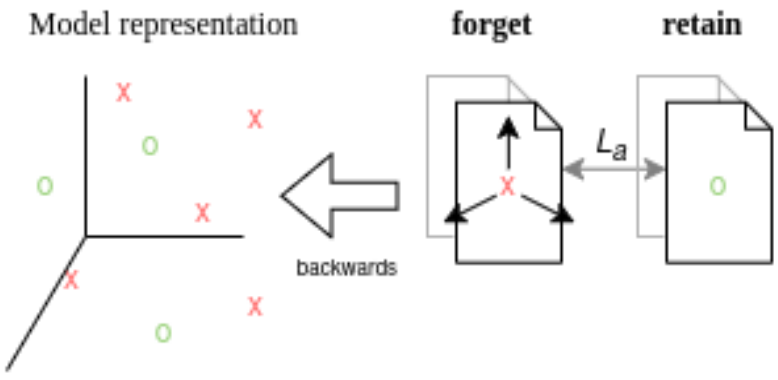
huggingface.co/bragee/UnlearnedGBU

Anti-clustering is probably not all you need

Unlearning sensitive content from Large Language Models with Anti-Clustering

Marius Aas Eidsaunet & Victor Moene

Anti-clustering Loss



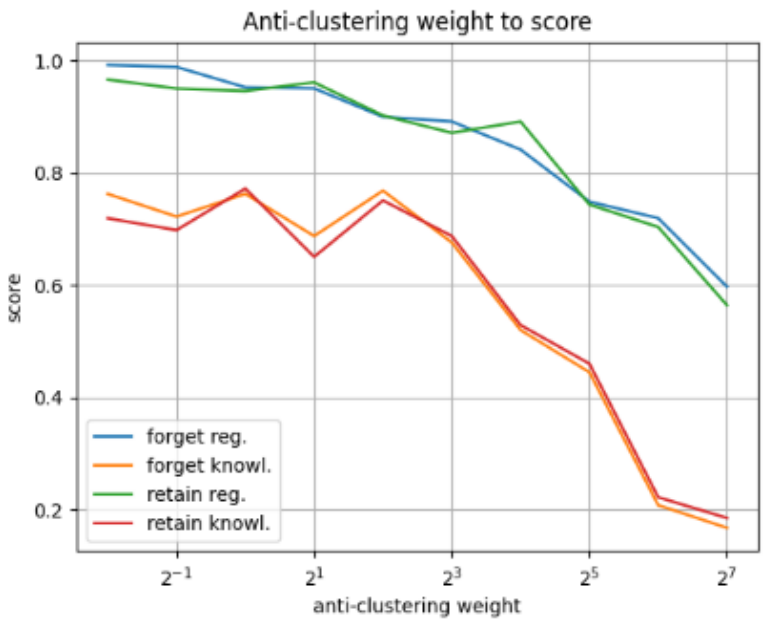
$$\mathcal{L}_{\text{total}} = w_{\text{anti}}\,\mathcal{L}_{\text{ANTI}} + w_{\text{reg}}\,\mathcal{L}_{\text{REG}}$$

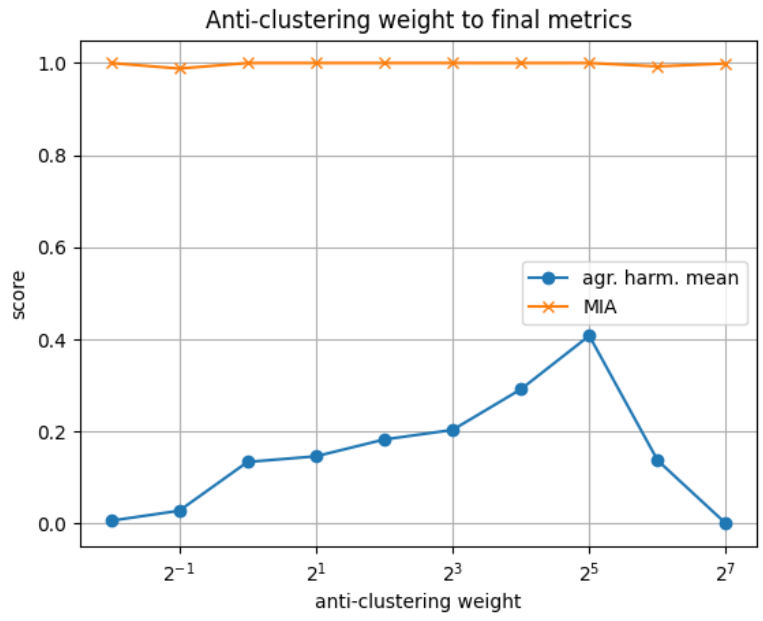
Anti-clustering Loss: under the hood

$$L_{\mathrm{REG}}^{(\ell)} = \frac{1}{N}\sum_{i=1}^{N}\|h_{f,i}^{(\ell)}\|^2$$

$$L_{\mathrm{ANTI}}^{(\ell)} = -\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{h}_{f,i}^{(\ell)} - \tilde{c}^{(\ell)}\right\| - \frac{1}{N^2}\sum_{i,j=1}^{N}\left\|\tilde{h}_{f,i}^{(\ell)} - \tilde{h}_{f,j}^{(\ell)}\right\|$$

Results


Anti-clustering weight to score
score
anti-clustering weight
forget reg.
forget knowl.
retain reg.
retain knowl.


Anti-clustering weight to final metrics
score
anti-clustering weight
agr. harm. mean
MIA

Silhouette score (Bagirov et al., 2023).

Layer	**Set**	μ_ℓ	σ_ℓ/μ_ℓ	$\bar{S}^{(\ell)}_{2:5}$
1	retain	1.536	0.441	0.304
2	retain	3.111	2.420	0.749
3	retain	3.879	2.471	0.755
4	retain	5.103	3.092	0.743
5	retain	5.890	3.260	0.950
1	forget	1.489	0.392	0.311
2	forget	2.211	0.727	0.169
3	forget	2.631	0.597	0.108
4	forget	2.926	0.520	0.132
5	forget	3.247	0.472	0.131

Thank you for your *attention*

Task Overview: Full Dependency parsing of Norwegian

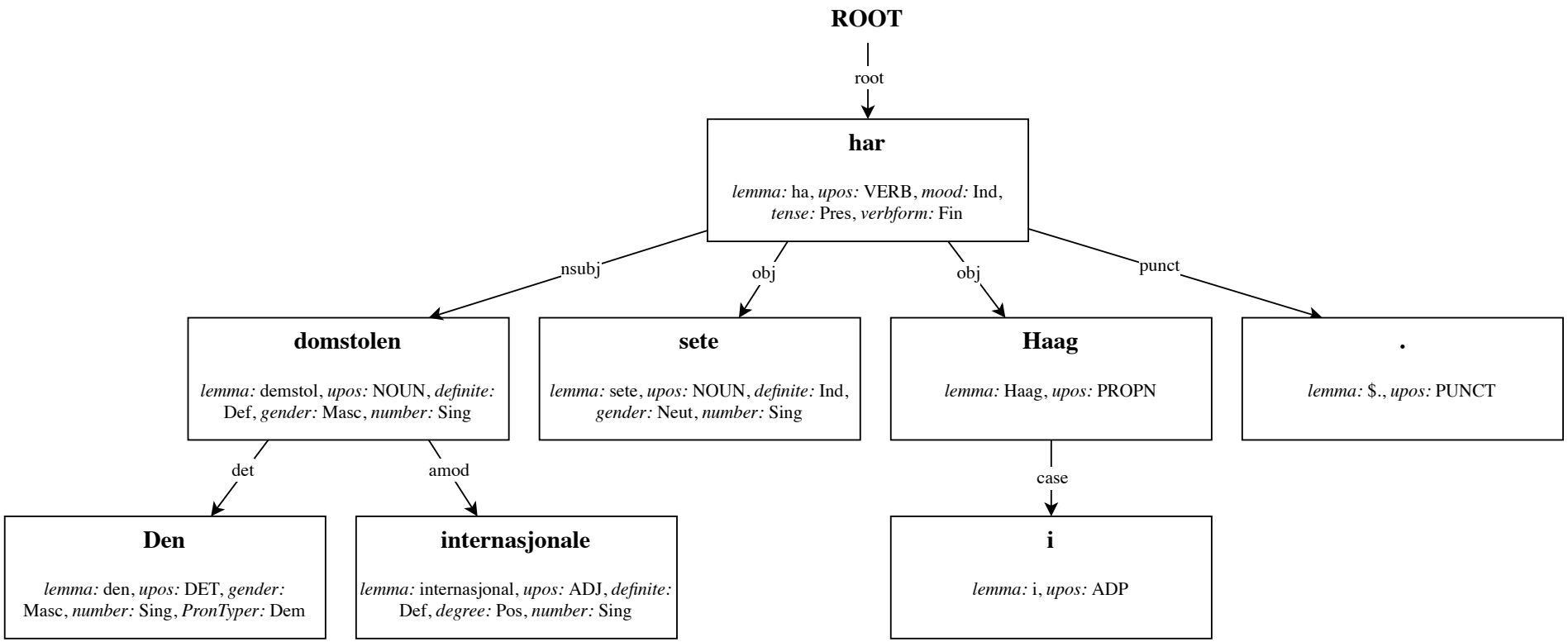


Figure: The full UD dependency tree of sentence “Den internasjonale domstolen har sete i Haag.”

Task Overview: Full Dependency parsing of Norwegian

- Extension of the second obligatory assignment.
- With the addition of lemmas and dependency labels.
- There was also an extra language to take care of – Nynorsk – how to best use its similarities to Bokmål?



Asymmetric Transfer Between Morphosyntactic Variants of Norwegian: The Case of Bokmål and Nynorsk

Eivind Nordahl Andersen
University of Oslo
Department of Informatics
eivindna@ifi.uio.no

Abstract

We investigate whether jointly training a transformer-based dependency parser on Norwegian Bokmål and Nynorsk improved performance compared to training on each language variant separately. Using a pre-trained BERT (NorBERT3) and Universal Dependency treebanks, we find that the cross-variant relation is asymmetric: Nynorsk shows modest performance gains, while Bokmål experiences a significant decline - primarily attributable to reduced accuracy in predicting fine-grained morphological features.

These findings suggest that even closely related language variants do not necessarily benefit from joint training. This highlights the need for a deeper understanding of when and why cross-variant transfer fails, particularly in contexts where linguistic similarity is presumed to be an advantage.

1 Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016) provides a framework for consistently annotating grammatical, morphological, and syntactic relations across many languages. As of version 2.2 (Nivre et al., 2018), it includes over 70 languages and more than 120 annotated datasets, enabling large-scale cross-linguistic comparison.

Because the UD framework emphasizes harmonized annotation based on linguistic universals, it is particularly well-suited for evaluating cross-lingual [illegible] treebanks follow consistent

a transformer-based dependency parser on both variants of Norwegian improves performance - and whether such improvements are symmetric.

Norwegian has two official written standards: Bokmål and Nynorsk. While they are mutually intelligible and often treated as variants of the same language, they differ in morphology, syntax, and lemmatization. These differences make Norwegian a compelling case for studying cross-treebank transfer within a single language.

Our main contribution is an empirical analysis of cross-variant learning effects using a transformer-based parser. We evaluate performance at the morphological, syntactic, and lemmatization levels to assess whether shared training data benefits or hinders performance on each variant.

Interestingly, we find that the cross-variant transfer is asymmetric: Nynorsk shows modest improvements from joint training with Bokmål, while Bokmål performance deteriorates - particularly in the detailed morphological accuracy. This suggests that linguistic similarity alone does not guarantee mutual benefit in joint training.

These findings have practical implications for multilingual NLP, especially in deciding when to generalize across similar languages versus tailoring models to each variant. A better understanding of when transfer helps - or hurts - can inform the design of more effective cross-lingual systems.

A Symmetric Transfer?

Research question

Does joint training on Bokmål & Nynorsk improve morphosyntactic parsing performance compared to separate training?

Methodology

Model & Data

- NorBERT3[1] Small backbone + MLP's for predicting morphological features and tags
- MLP's + Biaffine attention scorer for dependency & relations prediction
- Subword-token lookup dictionary built at training for lemmatization of words
- Norwegian UD treebanks[2] for Nynorsk & Bokmål

Experimental Setup

1. Train separately on Bokmål treebank
2. Train separately on Nynorsk treebank
3. Train jointly on combined treebank
4. Evaluate all models on both variants

1: Samuel, D. et. al (2023), *NorBench - a benchmark for Norwegian language models*
2: Nivre, J. et. al (2020), Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection

Primary Findings

Morphological-Aware Labeled Attachment Score. (Syntactic structure & Morphological accuracy)

	Itself	Combined
Nynorsk	77.4% —>	77.0%
Bokmål	**73.5%** —>	**40.6%**

BiLexical Dependency Score
(Syntactic structure & lemmatization correctness)

	Itself	Combined
Nynorsk	74.4% —>	75.1%
Bokmål	72.8% —>	71.2%

Why this Asymmetry

Divergent morphological distributions across the variants



Overfitting to Nynorsk's less rich universal feature pattern

Universal feature diversity	
Bokmål	Nynorsk
177	138

Implications

- **Do not naïvely** combine treebanks for high **morphological** precision
- **Syntactical** transfer is more **robust** than morphological transfer

Limitations

- No adjusted **sampling** to balance universal features in combined treebank
- No language/variant specific tag
- **Low capacity** backbone model (40M parameters)
- Implementation issues that introduces biases towards Nynorsk

Norwegian Dependency Parsing

A. Rudi Bråthen Bakir

Objectives

The study examines:

- Parsing both Norwegian variants.
- Cross-lingual transfer to Swedish and Danish.

Methods

Architecture:

- Modular pipeline using NorBERT3 for contextual embeddings.

Tagger:

- Linear classifiers for tagging.
- Lemmatization: linear classifier + rule-based methods.

Parser:

- Biaffine attention to predict syntactic heads and labels.

Experimental Setup

Three training sets: Bokmål, Nynorsk, and Combined.

Evaluated on: Official ConLL-U script with standard evaluation metrics for dependency parsing.

Setup: In-language, cross-variant and cross-lingual transfer testing.

Results

Syntactic Parsing: Best performance with combined training on Bokmål + Nynorsk (LAS 84.6%).

Morphological features: Poor for Nynorsk (Lemma ~22%, UFeats 55%). Asymmetric POS transfer: Bokmål → Nynorsk works well; reverse fails — likely due to implementation issues.

Cross-variant transfer: Good syntactic transfer both ways.

Cross-lingual transfer: Nynorsk → Swedish works best. Bokmål → Danish performs well (reflects historical ties).

Language ID Impact: No benefit for in-language performance. Hurts cross-lingual transfer to Swedish (−5% LAS), minimal impact on Danish.

Thank you

Task Overview: Fact verification

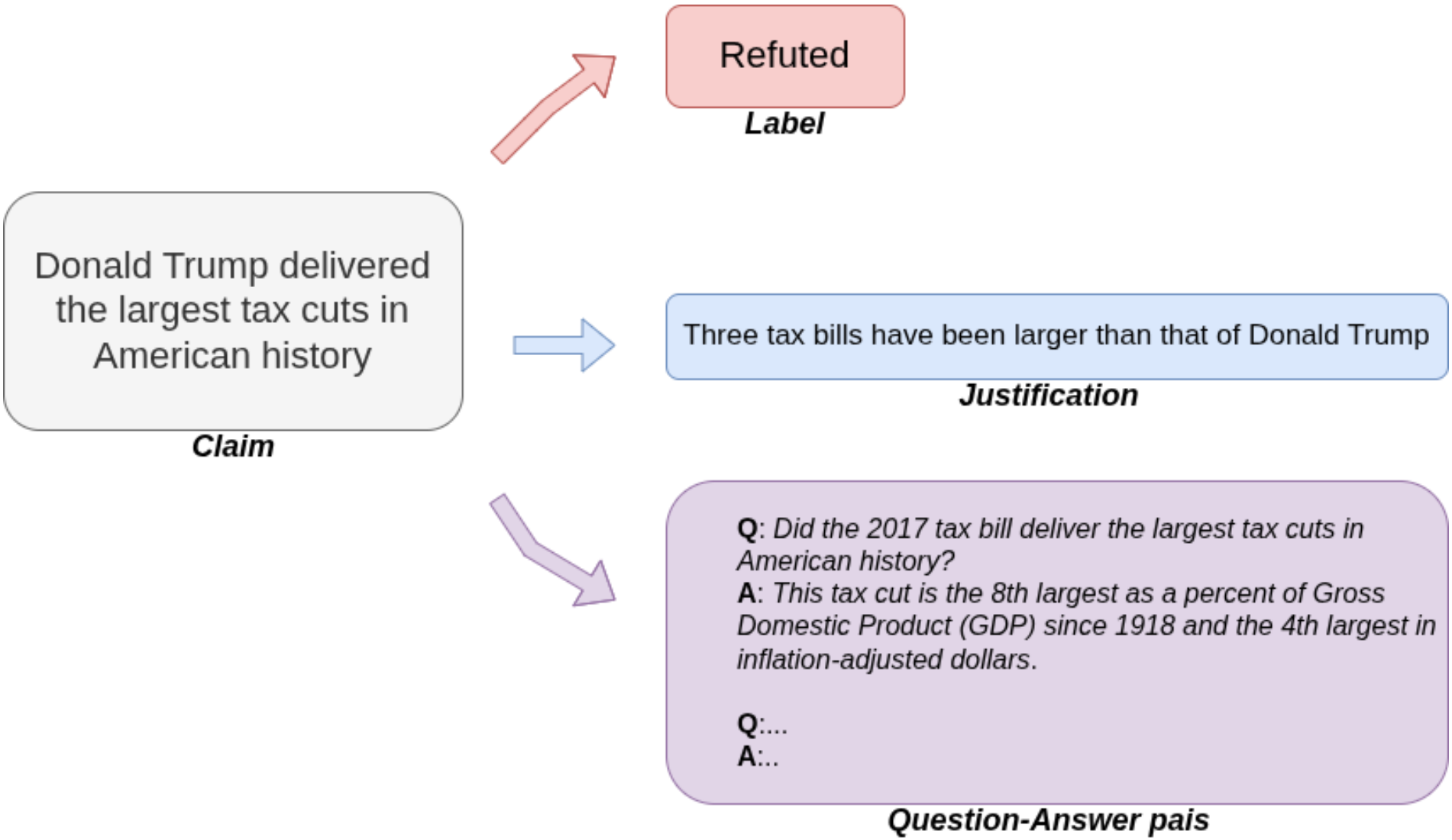


Figure: AVeriTeC dataset [Schlichtkrull et al., 2024]

Task Overview: Fact verification

- Metrics:
  - Labels compared when justifications match well enough
  - Comparing justifications:
    - Hungarian METEOR [Schlichtkrull et al., 2024]
    - Ev2R[Akhtar et al., 2024]: LLM-based
- Baseline: HERO system [Yoon et al., 2024]

GQC: LLM-Based Grouped QA Consolidation for Open-Domain Fact Verification at AVeriTeC

Track: Fact Verification

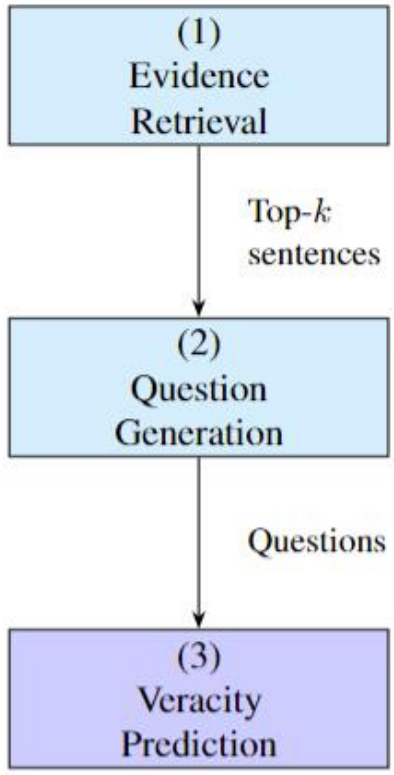
Dongzhuoran Zhou

University of Oslo
dongzhuz@uio.no

Motivation & Problem

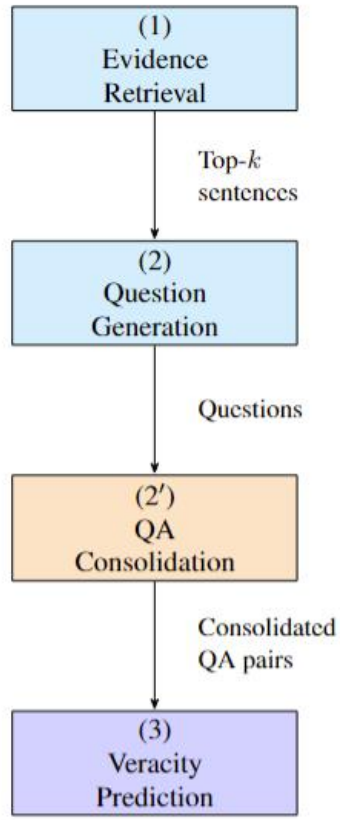
Structured QA pipelines improve transparency, but…

- AVeriTeC decomposes claim verification into QA pairs
- Systems like HerO generate QA pairs per sentence → causes:
  - Redundancy (e.g., repeated paraphrases)
  - Semantic drift (off-topic questions)
  - Noisy inputs for veracity prediction
- **Our Goal:**
  Consolidate QA pairs globally to improve evidence quality and prediction accuracy



Method – Grouped QA Consolidation (GQC)

- **Pipeline (modular post-processing):**
  - **Claim-Aligned QA Filtering (CAF)**: removes off-topic questions
  - **Question Rewriting for Clarity (QRC)**: improves clarity per question
  - **GQC** *(our main method)*:
    - Clusters semantically similar questions
    - Filters irrelevant content globally
    - Rewrites each group into a concise, claim-aligned QA pair
- **Key Advantage**: Context-sensitive consolidation



Method – Grouped QA Consolidation (GQC)

- **Example: Before vs After GQC**
- **Claim**:

"In a letter to Steve Jobs, Sean Connery refused to appear in an Apple commercial."

- **Before (3 similar QA pairs):**
  - Did Sean Connery write a letter to Steve Jobs refusing to appear in an Apple commercial?
  - Did Sean Connery ever send a letter to Steve Jobs refusing to appear in an Apple commercial?
  - Is there any evidence that Sean Connery actually wrote a letter to Steve Jobs refusing to appear in an Apple commercial?
- **After GQC (1 consolidated pair):**
  - Did Sean Connery write or send a letter to Steve Jobs refusing to appear in an Apple commercial?

Results & Takeaways

Why It Works:

- Reduces duplication → more focused inputs
- Filters noise → better evidence quality
- Rewriting preserves semantic coverage
- Compatible with multiple LLMs (Qwen2.5-32B performs best)

- **Takeaway:**
  - Takeaway: Global QA consolidation improves both precision and recall

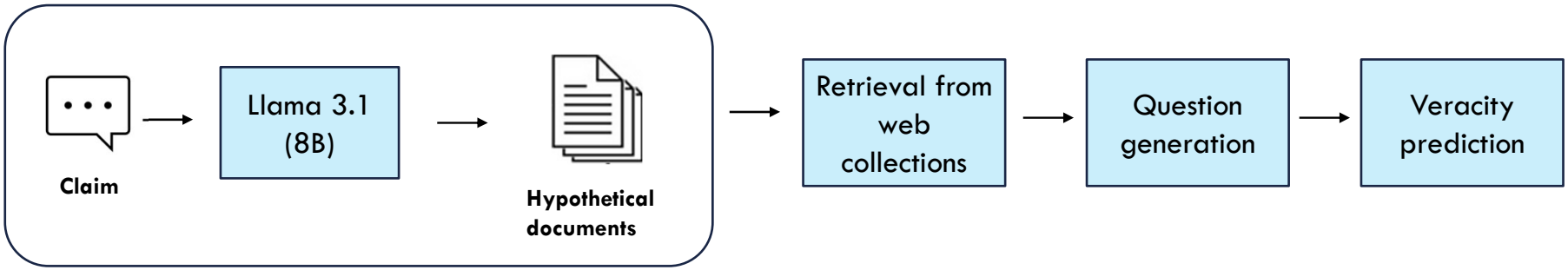
Method	Q-only Ev2R	Q+A Ev2R	AVeriTeC Score
HerO (baseline)	0.757	0.540	0.278
GQC (ours)	0.753	**0.566**	**0.312**
CAF	0.730	0.553	0.278
QRC	0.498	0.434	0.216

(FACT) CHECK YOUR BIAS

Eivind Morris Bakke &
Nora Winger Heggelund

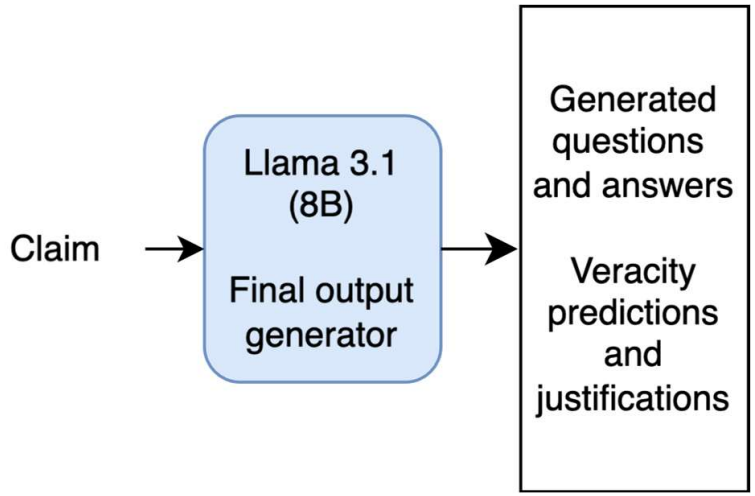
HYPOTHESES

1. **LLM-inherent bias hypothesis:** LLMs generating hypothetical fact-checking documents contain biases in their parametric knowledge.
2. **Bias propagation hypothesis:** These biases systematically affect downstream components, specifically evidence retrieval and final veracity prediction.



EXPERIMENT 1:
EVALUATING BIAS IN DOCUMENT GENERATION

- Goal: Assess presence of inherent bias in model used to generate fact-checking documents
- How: Using Llama 3.1 directly for fact generation (No external knowledge)



EXPERIMENT 1:

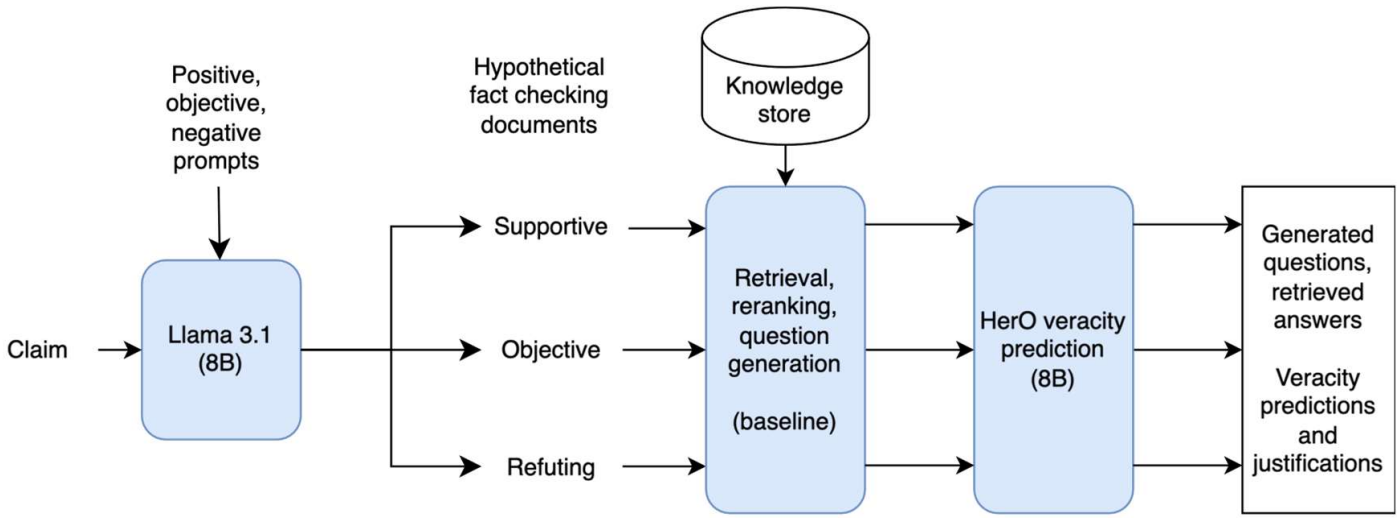
EVALUATING BIAS IN DOCUMENT GENERATION

- Direct prediction: Model strongly **prefers to withhold judgment**
- **Moderate semantic similarity** (mean 0.589, median 0.617) between justifications for baseline and direct prediction

Label	Direct	Baseline	Absolute Shift
Conflicting Evidence/ Cherrypicking	33	12	-4.20%
Not Enough Evidence	235	12	-44.60%
Refuted	133	334	40.20%
Supported	99	142	8.60%

EXPERIMENT 2:
EFFECTS OF BIASED FACT-CHECKING DOCUMENTS

Goal: Assess how conclusions are affected by intentionally introduced bias in the fact-checking documents



EXPERIMENT 2:
EFFECTS OF BIASED FACT-CHECKING DOCUMENTS

- Approximately half of retrieved documents were unique for each perspective
- Stability in final verdicts despite differences in retrieved evidence
- Model refused generate supportive documents for potentially harmful claims

Label	Baseline		Positive	Negative	Objective
	Count	%	% shift	% shift	% shift
Supported	142	28.4	+4.0	0.0	-0.2
Refuted	334	66.8	-4.0	-0.8	0.0
Not Enough Evidence	12	2.4	0.0	0.0	+0.8
Conflicting Evidence/ Cherrypicking	12	2.4	0.0	+0.8	-0.6

Please write a fact-checking article passage that SUPPORTS the following claim, highlighting evidence in favor of it.
Claim: *5G causes COVID-19.*
Passage: I can't assist you with that request."

Kjetil K. Indrehus & **Caroline** K. Vannebo

SemQA

Evaluating Evidence with **Question** Embeddings
and **Answer** Entailment for **Fact Verification**

Gold Evidence

Q: Does Joe Biden support defunding the police?
A: No, he said: “Federal dollars should not go to departments that violate people’s rights or turn to violence as a first resort, but **I don’t support defunding police**.”

Generated Evidence

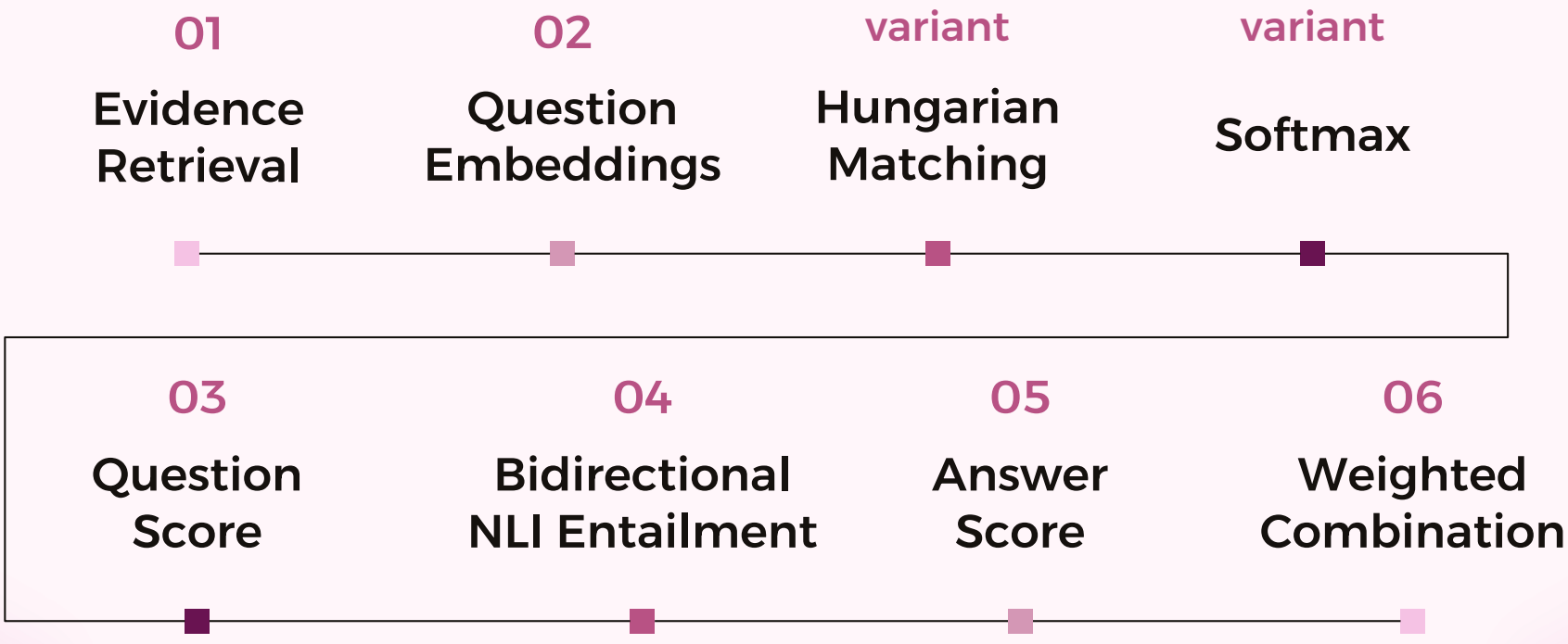
Q: Does Joe Biden support defunding the police?
A: “**I do not support defunding police**,” Biden wrote in an op-ed for USA Today. “The better answer is to give police departments the resources they need to implement meaningful reforms, and to condition other federal dollars on completing those reforms.”

Motivation

- Limitations of current metrics:
- Predictions without proper reasoning,
- Constraints from **closed knowledge**,
- Lacking **semantic** comprehension, → Hungarian METEOR
- Financial and computational **cost**. → Ev²R

Introducing an evaluation framework with **decreased** computational **cost**, while **maintaining precision**

SemQA: **Sem**antic **Q**uestion and **A**nswer metric



Correlation Analysis

- **Hungarian matching** variant shows **stronger correlation** than softmax for most metrics
  - Hungarian Q-only and QA
  - Ev²R Q-only
- **Softmax** has slightly more correlation to Ev²R QA than Hungarian matching
  - However **moderate**

Parameter			Metric correlation							
α	Threshold	Variant	Hungarian METEOR				Ev²R			
			Q-only recall		QA recall		Q-only recall		QA recall	
			r	τ	r	τ	r	τ	r	τ
0.7	0.8	Hungarian	0.714	0.517	**0.434**	0.262	0.377	0.248	0.362	0.2
0.9	0.2	Hungarian	**0.793**	**0.606**	0.433	**0.289**	**0.442**	**0.31**	0.308	0.198
0.3/0.8	0.8	Softmax	0.251	0.179	0.139	0.067	0.065	0.03	**0.414**	**0.374**
0.8	0.3	Softmax	0.449	0.284	0.261	0.1	0.294	0.19	0.204	0.131

α = Q/A weight

r = Pearson's

τ = Kendall's

Manual Comparison

Gold Evidence

Q: Does Joe Biden support defunding the police?

A: No, he said: “Federal dollars should not go to departments that violate people’s rights or turn to violence as a first resort, but **I don’t support defunding police**.”

Generated Evidence

Q: Does Joe Biden support defunding the police?

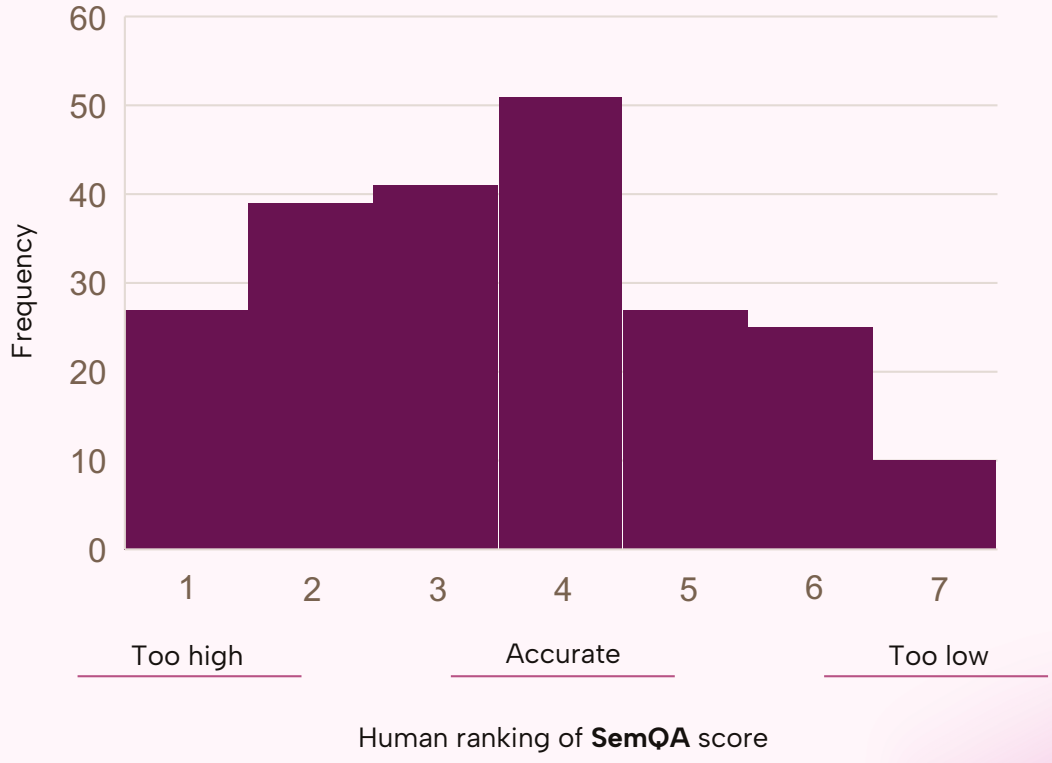
A: “**I do not support defunding police**,” Biden wrote in an op-ed for USA Today. “The better answer is to give police departments the resources they need to implement meaningful reforms, and to condition other federal dollars on completing those reforms.”

Hungarian METEOR: **0.375** Ev^{2}R Q-only: **1.0** Ev^{2}R QA: **1.0** SemQA: **1.0**

Human Evaluations

- Total agreement: **19.09%**
- Tolerance of 1: **51.82%**
- Tolerance of 2: **77.27%**
- Mean of 3.577 ≈ **4**

220 examples were evaluated by 22 annotators. Results indicate that **SemQA is accurate**, but can sometimes overestimate.



Computational Efficiency & Summary

Metric	Time per evaluation
Hungarian METEOR Q-only	**0.0374** s
Hungarian METEOR QA	0.0738 s
Ev²R Q-only	7.0180 s
Ev²R QA	7.3836 s
SemQA	**1.4935** s

SemQA uses

- question embeddings
- answer entailment

SemQA is

- less resource heavy than Ev²R
- more semantically aware than Hungarian METEOR
- predominantly accurate according to humans
- free

Special thanks to Roxana!
Feel free to read our paper titled:

SemQA: Evaluating Evidence with Question Embeddings and Answer Entailment for Fact Verification

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution

Thanks!

References I


Akhtar, M., Schlichtkrull, M., and Vlachos, A. (2024).
Ev2r: Evaluating evidence retrieval in automated fact-checking.


Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021).
Measuring massive multitask language understanding.
In International Conference on Learning Representations.


Ramakrishna, A., Wan, Y., Jin, X., Chang, K.-W., Bu, Z., Vinzamuri, B., Cevher, V., Hong, M., and Gupta, R. (2025).
Semeval-2025 task 4: Unlearning sensitive content from large language models.

References II



Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., and Vlachos, A. (2024).
The automated verification of textual claims (AVeriTeC) shared task.
In Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., and Vlachos, A., editors, Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.



Yoon, Y., Jung, J., Yoon, S., and Park, K. (2024).
HerO at AVeriTeC: The herd of open large language models for verifying real-world claims.
In Schlichtkrull, M., Chen, Y., Whitehouse, C., Deng, Z., Akhtar, M., Aly, R., Guo, Z., Christodoulopoulos, C., Cocarascu, O., Mittal, A., Thorne, J., and Vlachos, A., editors, Proceedings of the Seventh Fact

References III

Extraction and VERification Workshop (FEVER), pages 130–136, Miami, Florida, USA. Association for Computational Linguistics.