UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D22 – Final Report on Efficient Inference

Nature	Report	**Work Package**	WP6
Due Date	30/09/2025	**Submission Date**	29/09/2025
Main authors	Tsz Kin Lam (UEDIN)		
Co-authors			
Reviewers	André F. T. Martins (IT)		
Keywords	Efficient, inference, compact, evaluation, tools		
Version Control			
v0.1	**Status**	Draft	08/09/2025
v1.0	**Status**	Final	29/09/2025

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Introduction **5**

2 Task T6.1: Inference Speed (UEDIN) **7**

2.1 Output 1: Prepending or Cross-Attention for Speech-To-Text? An Empirical Comparison 7

3 Task T6.2: Model Compression (UEDIN) **8**

3.1 Output 1: ExpertSteer: Intervening in LLMs through Expert Knowledge 8

4 Task T6.3: Providing Tools for Evaluation, Usage, Sharing and Adaptation of the Models (UEDIN, UNB) **9**

4.1 Output 1: No Train but Gain: Language Arithmetic for training-free Language Adapters enhancement 9

4.2 Output 2: xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection 10

4.3 Output 3: Nimification of EuroLLM 11

5 Conclusion **13**

List of Figures

1 Representation of the architectures analyzed in the paper. Both (a) and (c) are based on encoder-decoder architecture but (a) uses cross-attention, whereas (c) uses DFP. Secondly, both (b) and (c) uses DFP, but (c) contains a speech encoder, making it not decoder-only. The (audio) causal masking can be applied to both the previous tokens and the audio sequence or only to the previous tokens. 7

2 An overview of EXPERTSTEER, including four steps: (1) aligning the dimensionality of the expert and target models, (2) identifying the layer pairs to be intervened upon, (3) generating steering vectors from the expert model, and (4) intervening in the generation process of the target model. 9

3 Language arithmetic as an extension of the MAD-X framework. Given language and task adapters (left), language arithmetic (right) enables post-processing, training-free improvement in two use-cases: (i) zero-shot where a language adapter for a target language was not trained (presented in the figure as Spanish, which was not part of existing language adapters pool, LA_{es}(en, fr)) or (ii) to improve existing language adapters via arithmetic with either related language or a language on which task adapter was trained (e.g. LA_{fr}(en, fr)). 10

4 The xCOMET framework illustrated through a real example: the metric not only provides a sentence-level score, but also predicts translation error spans along with their respective severity. From these spans, we can infer MQM score (following the MQM typology) that informs and highly correlates with the sentence-level score (see Section 6). These spans complement the sentence-level score by providing a detailed view into the translation errors. 11

5 Partnership between NVIDIA and UTTER on the nimification of EuroLLM. 12

Abstract

This deliverable presents the final outcomes of WP6, covering (T6.1) efficient inference, (T6.2) model compression, and (T6.3) tools evaluation, usage, sharing and adaptation of the models.

In the previous deliverable, we reported our completed work about model distillation, lexical shortlisting and multimodal model compression using discreet speech units. Building on this foundation, we continue to advance methods for efficient inference across these tasks. In task T6.1, we extend our caching-and-distilling approach by introducing an uncertainty-based selection method to further optimize LLM calls. For multimodal models, we conducted a a systematic comparison of cross-attention and prepending in encoder-decoder architectures with continuous speech inputs. In task T6.2, we have explored activation steering, a new technique that yields significant performance improvements while preserving inference efficiency. In task T6.3, on the evaluation side, we continue to innovate by completing a new machine translation evaluation toolkit called `xCOMET`, which integrates fine-grained errors. Finally, in collaboration with NVIDIA, we optimized the EuroLLM model for enterprise AI workloads.

In total, our contribution includes three conference papers, one journal article, one arXiv preprint and an external industry collaboration.

1 Introduction

Proposal

This WP addresses the call text "proposed solutions should be energy efficient" by optimising model inference. Efficient inference is critical to applications. In our customer support and online meeting assistant scenarios, latency is critical to the user experience. Beyond the immediate proposal, we anticipate that third parties will benefit from reduced computational requirements to build models on custom data or integrate inference into their applications. A second order benefit of efficiency is improved privacy because it supports application developers running inference locally rather than on the cloud.

WP6 has the following main goals: reduce inference latency and throughput; decrease the size of the model. Many methods trade between efficiency and quality of the model. There is no single speed cutoff nor a single acceptable level of quality loss, but rather a range of options depending on application requirements. For each of these goals, our aim is to push the Pareto frontier i.e. with faster speed but the same quality loss or better quality for the same speed.

Work Presented in the First Deliverable

In the first deliverable, we first completed a survey paper on efficiency (Treviso et al., 2023) which guides the development of the subsequent works. We made an initial contribution to the hard problem of lexical shortlisting for large language models (Bogoychev et al., 2024) and looked at more efficient use of LLMs by caching and distilling the responses from a (teacher) Large Language Model (LLM) (Ramírez et al., 2024). We also explored effective and efficient question-answer representations (Hu et al., 2024). In T6.2 on model compression, we started work on smaller models for multimodal speech translation using discrete speech units (Lam et al., 2024). Finally for the third task on tools for evaluation, usage and sharing, we delivered `TowerEval` (Alves et al., 2024) which is an evaluation framework for large language models.

In total, we have completed four conference papers on Task 6.1, one paper on Task 6.2 and one paper on T6.3. At that time, the discrete-speech-units paper and the caching-and-distilling paper were under review and have now been accepted to IWSLT2024 and ACL (Findings) 2024, respectively.

Work Presented in this Deliverable

In task T6.1, we continued research about optimizing LLM calls by introducing an uncertainty-based selection method (Ramírez et al., 2024). In multimodal language model, we conducted a systematic comparison of cross-attention and prepending in encoder-decoder architectures using continuous speech representation as inputs (Lam et al., 2025). Our results reveal better efficiency in terms of generation speed and GPU memory footprint of cross-attention over prepending in integrating speech into language model. In task T6.2, we have explored activation steering, a new technique to control LLM's generation. Our distillation-based modification brings remarkable performance gain over the existing methods when applied to smaller language models (Wang et al., 2025). In the task T6.3, we leveraged language arithmetic for efficient cross-lingual adaptation of LLMs (Klimaszewski et al., 2025). We also completed a new machine translation evaluation toolkit called `xCOMET` (Guerreiro et al., 2024), which integrates fine grained errors. Finally, we partnered with NVIDIA to "nimify" our EuroLLM model for enterprise AI workloads.

Our achievement is summarized as follows:

Summary of Output

Manuscripts:

- 1 journal article (TACL24)
- 1 arXiv pre-print
- 2 conference papers (NAACL25, COLING25)

Code and data:

- NAACL25: https://github.com/hlt-mt/FBK-fairseq/
- COLING25: https://github.com/mklimasz/language-arithmetic
- xCOMET: https://huggingface.co/Unbabel/XCOMET-XL

Events: NVIDIA at VivaTech 2025 (June 11-14 — Paris, France) https://www.nvidia.com/en-eu/gtc/vivatech/

List of completed works and their related sections:

Task	Section	Paper
	T6.1 §2.1	Tsz Kin Lam, Marco Gaido, Sara Papi, Luisa Bentivogli, Barry Haddow "Prepending or Cross-Attention for Speech-to-Text? An Empirical Comparison", NAACL 2025
T6.2	§3.1	Weixuan Wang, Minghao Wu, Barry Haddow, Alexandra Birch "EXPERTSTEER: Intervening in LLMs through Expert Knowledge", Under Review
T6.3	§4.1	Mateusz Klimaszewski, Piotr Andruszkiewicz, Alexandra Birch "No Train but Gain: Language Arithmetic for training-free Language Adapters enhancement", COLING 2025
	§4.2	Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, André F. T. Martins "xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection", TACL 2024
	§4.3	Nimification of EuroLLM

Table 1: List of publications to be discussed

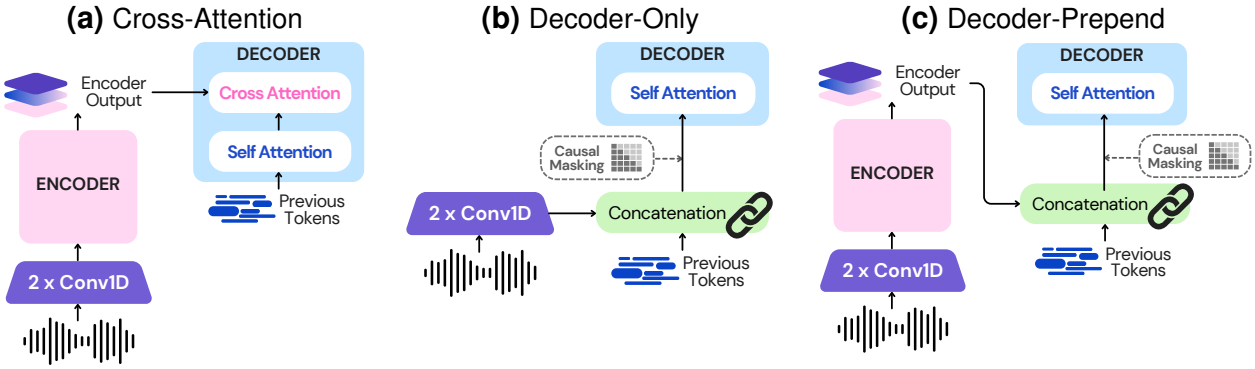

Figure 1: Representation of the architectures analyzed in the paper. Both (a) and (c) are based on encoder-decoder architecture but (a) uses cross-attention, whereas (c) uses DFP. Secondly, both (b) and (c) uses DFP, but (c) contains a speech encoder, making it not decoder-only. The (audio) causal masking can be applied to both the previous tokens and the audio sequence or only to the previous tokens.

2 Task T6.1: Inference Speed (UEDIN)

Proposal highlights

- Cross-attention or prepending in speech-to-text language modeling

2.1 Output 1: Prepending or Cross-Attention for Speech-To-Text? An Empirical Comparison

Following the remarkable success of Large Language Models (LLMs) in NLP tasks, there is increasing interest in extending their capabilities to speech—the most common form of communication. The most widespread approach to integrating speech into LLMs is dense feature prepending (DFP), which prepends the projected speech representations to the textual representations, allowing end-to-end training with a speech encoder. This raises questions about the need for a sophisticated speech encoder for DFP and how its performance compares with a cross-attention based encoder-decoder architecture.

Data and models. To perform a controlled architectural comparison, we train all models from scratch rather than using large pretrained models and use comparable data and parameter settings, testing speech-to-text recognition (ASR) and translation (ST) on MuST-C v1.0 (Di Gangi et al., 2019) and CoVoST2 (Wang et al., 2021) datasets in both Transformer and Conformer (Gulati et al., 2020) speech encoders.

Methodology. We compare DFP and cross-attention (see Fig. 1) under a variety of configurations, such as CTC compression (Liu et al., 2020; Gaido et al., 2021), sequence-level knowledge distillation (Kim and Rush, 2016), on monolingual, bilingual, and multilingual models.

Findings. Despite the wide adoption of DFP, our results do not indicate a clear advantage of DFP over cross-attention in both ASR and ST qualities, especially with a stronger speech encoder.

In contrast, cross-attention is indeed more efficient in terms of generation speed and GPU memory footprint.

3 Task T6.2: Model Compression (UEDIN)

Proposal highlights

- Transferrable activation steering vectors for smaller language models

Summary of completed work

In this task, we explored a new exciting area about model compression of LLMs via activation steering. Our proposed method extracts steering vector from huge LLMs, which could be applied to much smaller sized language models. Our results show substantial improvement over the existing methods in 15 popular benchmarks in four distinct domains.

3.1 Output 1: ExpertSteer: Intervening in LLMs through Expert Knowledge

Large Language Models (LLMs) exhibit remarkable capabilities across various tasks, yet guiding them to follow desired behaviours during inference remains a significant challenge. Activation steering offers a promising method to control the generation process of LLMs by modifying their internal activations. However, existing methods commonly intervene in the model's behaviour using steering vectors generated by the model itself, which constrains their effectiveness to that specific model and excludes the possibility of leveraging powerful external expert models for steering. To address these limitations, we propose EXPERTSTEER, a novel approach that leverages arbitrary specialized expert models to generate steering vectors, enabling intervention in any LLMs.

Data and models. We conduct comprehensive experiments using `Llama3.1-8B-Instruct` (Dubey et al., 2024), `Qwen2.5-7B-Instruct` (Yang et al., 2024) and `Gemma2-2B-Instruct` (Team et al., 2024) on 15 popular benchmarks across four distinct domains: (1) `medical` — medQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) and MMLU-Medical (Hendrycks et al., 2020), (2) `financial` — FPB (Malo et al., 2014), Flare-cfa (Xie et al., 2023), MMLU-Financial, (3) `mathematical` — GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021), MMLU-Math and (4) `general` — COPA (Roemmele et al., 2011), NLI (Conneau et al., 2018), ARC-C (Bhakthavatsalam et al., 2021), MMLU-Humanities, Salad (Li et al., 2024), Harmful Behaviors (ZySec-AI, 2024).

Methodology. EXPERTSTEER (Fig.2) transfers the knowledge from an expert model to a target LLM through a cohesive four-step process: (1) first aligning representation dimensions with autoencoders to enable cross-model transfer, (2) then identifying intervention layer pairs based on mutual information analysis, (3) next generating steering vectors from the expert model using Recursive Feature Machines (Radhakrishnan et al., 2024), and (4) finally applying these vectors on the identified layers during inference to selectively guide the target LLM without updating model parameters.

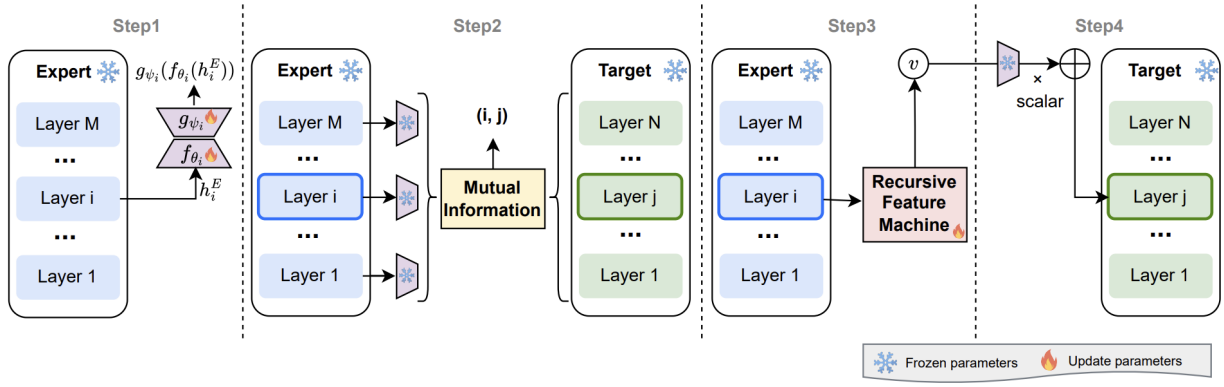

Figure 2: An overview of EXPERTSTEER, including four steps: (1) aligning the dimensionality of the expert and target models, (2) identifying the layer pairs to be intervened upon, (3) generating steering vectors from the expert model, and (4) intervening in the generation process of the target model.

Findings. Experiments demonstrate that EXPERTSTEER significantly outperforms established baselines, such as Supervised-Fine-Tuning (Wei et al., 2021), Knowledge Distillation (Boizard et al., 2024), and the SOTA steering baselines, including Inference-Time Intervention (Li et al., 2023), Contrastive Activation Addition (Panickssery et al.), and Semantic-Adaptive Dynamic Intervention (Wang et al., 2024) across diverse tasks at minimal cost.

4 Task T6.3: Providing Tools for Evaluation, Usage, Sharing and Adaptation of the Models (UEDIN, UNB)

Proposal highlights

- Language arithmetic for cross-lingual transfer in LLMs
- xCOMET: Transparent machine translation evaluation
- Nimification of EuroLLM for enterprise AI workloads

Summary of completed work

In this task, we completed an accepted work by COLING2025 on cross-lingual adaptation using language arithmetic. Following `TowerEval` in the last deliverable, we released `xCOMET`, a toolkit for evaluation of machine translation. Furthermore, we have collaborated with NVIDIA to optimize EuroLLM (Martins et al., 2025) for enterprise AI workloads.

4.1 Output 1: No Train but Gain: Language Arithmetic for training-free Language Adapters enhancement

Modular deep learning is the state-of-the-art solution for lifting the curse of multilinguality, preventing the impact of negative interference and enabling cross-lingual performance in Multilingual

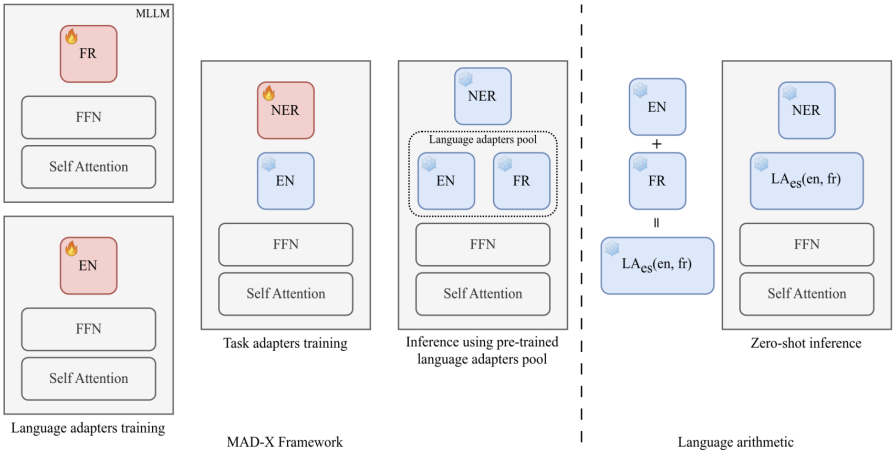

Figure 3: Language arithmetic as an extension of the MAD-X framework. Given language and task adapters (left), language arithmetic (right) enables post-processing, training-free improvement in two use-cases: (i) zero-shot where a language adapter for a target language was not trained (presented in the figure as Spanish, which was not part of existing language adapters pool, LA_{es}(en, fr)) or (ii) to improve existing language adapters via arithmetic with either related language or a language on which task adapter was trained (e.g. LA_{fr}(en, fr)).

Pre-trained Language Models. However, a trade-off of this approach is the reduction in positive transfer learning from closely related languages. In response, we introduce a novel method called language arithmetic, which enables training-free post-processing to address this limitation.

Data and models. The effectiveness of the proposed solution is demonstrated on three downstream tasks in a MAD-X-based (Pfeiffer et al., 2020) set of crosslingual schemes, acting as a post-processing procedure.

Methodology. Extending the task arithmetic framework, we apply learning via addition to the language adapters, transitioning the framework from a multi-task to a multilingual setup.

Findings. Language arithmetic consistently improves the baselines with significant gains, especially in the most challenging case of zeroshot application.

4.2 Output 2: xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection

Widely used learned metrics for machine translation evaluation, such as COMET and BLEURT, estimate the quality of a translation hypothesis by providing a single sentence-level score. As such, they offer little insight into translation errors (e.g., what are the errors and what is their severity). On the other hand, generative large language models (LLMs) are amplifying the adoption of more granular strategies to evaluation, attempting to detail and categorize translation errors. In this work, we introduce xCOMET, an open-source learned metric designed to bridge the gap between these approaches.

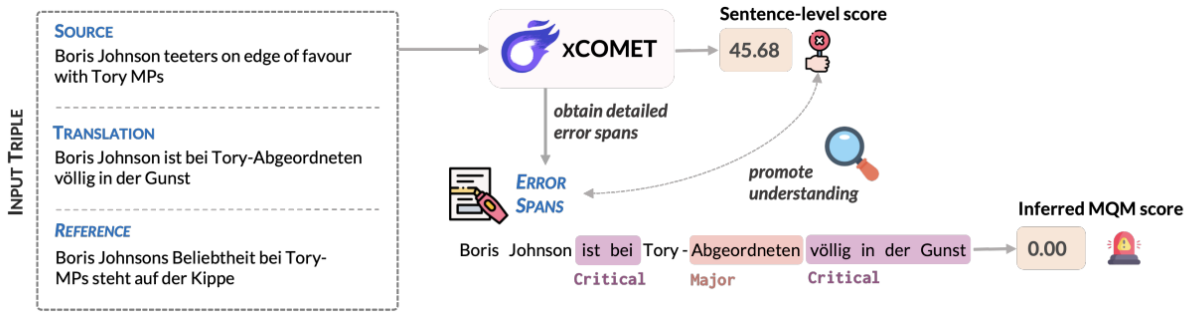

Figure 4: The xCOMET framework illustrated through a real example: the metric not only provides a sentence-level score, but also predicts translation error spans along with their respective severity. From these spans, we can infer MQM score (following the MQM typology) that informs and highly correlates with the sentence-level score (see Section 6). These spans complement the sentence-level score by providing a detailed view into the translation errors.

Data and models. We used both direct assessment (DA) and Multidimensional Quality Metric (MQM) data to train `xCOMET`. The DA annotations were collected by WMT from 2017 to 2020, and the MLQE-PE dataset (Fomicheva et al., 2022). As the MLQE-PE dataset does not contain reference translations, we used the post-edit translations as reference translations. Overall, the corpus consists of around 1 million samples, spanning 36 language pairs. The MQM annotations were sourced from WMT from 2020 to 2022.6 We also used annotations sourced from other MQ-Mannotated datasets: (i) IndicMT (Sai B et al., 2023), which contains MQM annotations spanning 5 Indian languages, and (ii) DEMETR (Karpinska et al., 2022), a diagnostic dataset with perturbations spanning semantic, syntactic, and morphological error categories.

We follow the same architecture of the scaled-up version of COMETKIWI detailed in (Rei et al., 2023), which uses a large pre-trained encoder model as its backbone encoder model.

Methodology. `xCOMET` integrates both sentence-level evaluation and error span detection capabilities, exhibiting state-of-the-art performance across all types of evaluation (sentence-level, system-level, and error span detection). Moreover, it does so while highlighting and categorizing error spans, thus enriching the quality assessment.

Findings. Our robustness analysis with stress tests show that `xCOMET` is largely capable of identifying localized critical errors and hallucinations. The model checkpoints have been available for the public[1].

4.3 Output 3: Nimification of EuroLLM

NVIDIA and UTTER partnered to optimize the EuroLLM model using NVIDIA Nemotron techniques to maximize cost efficiency and accuracy for enterprise AI workloads, including agentic AI. This exciting announcement was unveiled by NVIDIA CEO Jensen Huang at NVIDIA GTC Paris, VivaTech 2025. The full press release is shown in Fig.5, and the model is now available on https://build.nvidia.com/utter-project/eurollm-9b-instruct.

[1] https://huggingface.co/Unbabel/XCOMET-XL

NVIDIA

NVIDIA Partners With Europe Model Builders and Cloud Providers to Accelerate Region's Leap Into AI

- ***Model Builders Across Europe — Including France, Italy, Poland, Spain and Sweden — to Deliver Sovereign Models With NVIDIA Nemotron***
- ***AI Models Tailored to Local Languages and Culture Coming to Perplexity, Delivered as NVIDIA NIM Microservices and Hosted on Regional AI Infrastructure From NVIDIA Cloud Partners***

NVIDIA GTC Paris at VivaTech—NVIDIA today announced that it is teaming with model builders and cloud providers across Europe and the Middle East to optimize sovereign large language models (LLMs), providing a springboard to accelerate enterprise AI adoption for the region's industries.

Model builders and AI consortiums Barcelona Supercomputing Center (BSC), Bielik.AI, Dicta, H Company, Domyn, LightOn, the National Academic Infrastructure for Supercomputing in Sweden (NAISS) together with KBLab at the National Library of Sweden, the Slovak Republic, the Technology Innovation Institute (TII), the University College of London, the University of Ljubljana and UTTER are teaming with NVIDIA to optimize their models with NVIDIA Nemotron™ techniques to maximize cost efficiency and accuracy for enterprise AI workloads, including agentic AI.

Model post-training and inference will run on AI infrastructure in Europe from NVIDIA Cloud Partners (NCPs) participating in the NVIDIA DGX Cloud Lepton™ marketplace.

The open, sovereign models will provide a foundation for an integrated regional AI ecosystem that reflects local languages and culture. Europe's enterprises will be able to run the models on Perplexity, an AI-powered answer engine used to answer over 150 million questions per week. Companies will also be able to fine-tune the sovereign models on local NCP infrastructure through a new Hugging Face integration with DGX Cloud Lepton.

"Europe's diversity is its superpower — an engine of creativity and innovation," said Jensen Huang, founder and CEO of NVIDIA. "Together with Europe's model builders and cloud providers, we're building an AI ecosystem where intelligence is developed and served locally to provide a foundation for Europe to thrive in the age of AI — transforming every industry across the region."

Optimizing Model Accuracy and Inference Savings With NVIDIA Nemotron
Europe — the world's third largest economic region — is home to industries spanning manufacturing, robotics, healthcare and pharmaceuticals, finance, energy and creative.

To accelerate the region's AI-driven transformation, NVIDIA partners are delivering their open LLMs with support for Europe's 24 official languages. Several models also specialize in national language and culture, such as those from H Company and LightOn in France, Dicta in Israel, Domyn in Italy, Bielik.AI in Poland, the University of Ljubljana and the Slovak Republic models, BSC in Spain, NAISS and KBLab in Sweden, TII in the United Arab Emirates and the University College London in the U.K.

The LLMs will be distilled with NVIDIA Nemotron model-building techniques — including neural architecture search — as well as reinforcement learning and post-training with NVIDIA-curated synthetic data. These optimizations will reduce operational costs and boost user experiences by generating tokens faster during inference. The Nemotron post-training workloads will run on DGX Cloud Lepton hosted by European NCPs including Nebius, Nscale and Fluidstack.

Developers will be able to deploy the sovereign models as NVIDIA NIM™ microservices running on AI factories — on premises and across cloud service provider platforms — using a new NIM microservice that supports more than 100,000 public, private and domain-specialized LLMs hosted on Hugging Face.

Adding Europe's Sovereign AI Insights to Perplexity
Supporting AI diversity for enterprises across the region, Perplexity will integrate the sovereign AI models into its answer engine, which is used by European enterprises, publishers and organizations, including telecommunications and media giants. Perplexity uses LLMs to improve accuracy in search queries and AI outputs. The answer engine draws from credible sources in real time to accurately answer questions with in-line citations, perform deep research and complete assistive tasks.

"Perplexity's goal is to provide accurate, trustworthy answers to any question from any person, wherever they are," said Aravind Srinivas, cofounder and CEO of Perplexity. "Bringing NVIDIA-optimized sovereign AI models to Perplexity empowers innovation in Europe with AI built and running in the region."

Figure 5: Partnership between NVIDIA and UTTER on the nimification of EuroLLM.

5 Conclusion

This deliverable presents a summary of the most recent completed work on WP6 efficient inference. Our recent contribution includes four publications accepted to high-ranked NLP venues and one arXiv pre-print that is expected to appear on a major Machine Learning conference. Including the previous deliverable, this WP has led to nine publications. More importantly, UTTER made an successful collaboration with NVIDIA to promote EuroLLM for enterprise usage.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=EHPns3hVkj.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. Think you have solved direct-answer question answering? try arc-da, the direct-answer ai2 reasoning challenge. *arXiv preprint arXiv:2102.03315*, 2021.

Nikolay Bogoychev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. The ups and downs of large language model inference with vocabulary trimming by language heuristics. In Shabnam Tafreshi, Arjun Akula, João Sedoc, Aleksandr Drozd, Anna Rogers, and Anna Rumshisky, editors, *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 148–153, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.insights-1.17. URL https://aclanthology.org/2024.insights-1.17/.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *arXiv preprint arXiv:2402.12030*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://aclanthology.org/N19-1202/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. MLQE-PE: A multilingual quality estimation and post-editing dataset. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.530/.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. CTC-based compression for direct speech translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.57. URL https://aclanthology.org/2021.eacl-main.57/.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Zhanghao Hu, Yijun Yang, Junjie Xu, Yifu Qiu, and Pinzhen Chen. EEE-QA: Exploring effective and efficient question-answer representations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5520–5525, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.490/.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.649. URL https://aclanthology.org/2022.emnlp-main.649/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Mateusz Klimaszewski, Piotr Andruszkiewicz, and Alexandra Birch. No train but gain: Language arithmetic for training-free language adapters enhancement. In Owen Rambow, Leo Wanner,

Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11121–11134, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.737/.

Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Compact speech translation models via discrete speech units pretraining. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 114–124, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.16. URL https://aclanthology.org/2024.iwslt-1.16/.

Tsz Kin Lam, Marco Gaido, Sara Papi, Luisa Bentivogli, and Barry Haddow. Prepending or cross-attention for speech-to-text? an empirical comparison. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2994–3006, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.153. URL https://aclanthology.org/2025.naacl-long.153/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*, 2020.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. *URL https://arxiv. org/abs/2312.06681*, 3.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Bonnie Webber, Trevor Cohn, Yulan He,

and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617/.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

Guillem Ramírez, Alexandra Birch, and Ivan Titov. Optimising calls to large language models with uncertainty-based two-tier selection. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=T9cOYH0wGF.

Guillem Ramírez, Matthias Lindemann, Alexandra Birch, and Ivan Titov. Cache & distil: Optimising API calls to large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11838–11853, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.704. URL https://aclanthology.org/2024.findings-acl.704/.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73/.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95, 2011.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.795. URL https://aclanthology.org/2023.acl-long.795/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860, 2023. doi: 10.1162/tacl_a_00577. URL https://aclanthology.org/2023.tacl-1.48/.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. Covost 2 and massively multilingual speech translation. In *Interspeech 2021*, pages 2247–2251, 2021. doi: 10.21437/Interspeech.2021-2027.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*, 2024.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Expertsteer: Intervening in llms through expert knowledge. *arXiv preprint arXiv:2505.12313*, 2025.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

ZySec-AI. Harmful behaviors. 2024.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D22 Final Report on Efficient Inference