UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action

Number: 101070631

D20/D5.2 – Final Report on Uncertainty-Aware, Robust and Explainable Models

Nature	Report	**Work Package**	WP5
Due Date	30/09/2025	**Submission Date**	30/09/2025
Main authors	Wilker Aziz (UVA)		
Co-authors	Alexandra Birch (UEDIN), Barry Haddow (UEDIN), Bryan Eikema (UVA), Chrysoula Zerva (IT), Leonardo Ranaldi (UEDIN), Pierre Erbacher (NAV)		
Reviewers	Laurent Besacier (NAV)		
Keywords	uncertainty quantification, input attribution, hallucination, robustness		
Version Control			
v0.1	**Status**	Draft	26/09/2025
v1.0	**Status**	Final	30/09/2025

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

List of Figures

Abstract

In this report, we document WP5's outcomes throughout the entire project. We reiterate a summary of the output from RP1, and only describe in detail the output from RP2. At the end of the report, we also reflect briefly on impact.

1 Introduction

WP5 is focused on developing reliable and trustworthy underlying ML components for the core language technologies developed within UTTER. These components account for three themes:

- Uncertainty representation and estimation techniques for confidence-aware, self-critical AI assistants;
- Methods for explanation and attribution generation across domains and applications;
- Strategies to enhance robustness to noisy input.

These correspond to our three tasks, respectively, which we cover in detail in sections 2, 3, and 4.

Summary of Output

Manuscripts: 41 manuscripts (20 in RP1, and 21 in RP2); in RP2 we produced 10 conference papers (*ACL, EMNLP, COLM, AISTATS), 3 findings papers (*ACL and EMNLP), 2 workshop papers (UncertaiNLP), 2 journal articles (TACL), 4 preprints.
See Tables 1, 2, and 3 for outputs from each task.

Events:

- The First Workshop on Uncertainty-Aware NLP (collocated with EACL 2024, in RP1); Vázquez et al. (2024).
- The Second Workshop on Uncertainty-Aware NLP (collocated with EMNLP 2025) https://uncertainlp.github.io

Period	Venue	Paper	Code
RP1	EACL	Baan et al. (2024)	NA
	EACL	Ilia and Aziz (2024a)	https://github.com/evgeniael/predict_next_word
	EMNLP	Giulianelli et al. (2023)	https://github.com/dmg-illc/nlg-uncertainty-probes
	UncertaiNLP	Eikema (2024)	NA
	EACL	Waldendorf et al. (2024)	NA
	NAACL	Wang et al. (2024b)	https://github.com/weixuan-wang123/MONITOR
	ACL	Wang et al. (2024a)	https://github.com/weixuan-wang123/ReMaKE
	EMNLP	Zerva et al. (2022)	https://github.com/deep-spin/uncertainties_MT_eval
	EACL Findings	Ulmer et al. (2024)	https://github.com/Kaleidophon/non-exchangeable-conformal-language-generation
	ICLR	Farinhas et al. (2024)	https://github.com/deep-spin/non-exchangeable-crc
RP2	EMNLP	Eikema et al. (2025b)	https://github.com/roxot/structure-conditional-mbr
	COLM	Troshin et al. (2025a)	https://github.com/serjtroshin/selective_sampling
	UncertaiNLP	Troshin et al. (2025b)	NA
	UncertaiNLP	Groot et al. (2025)	https://github.com/tgroot56/Learning-to-vary-Teaching-LMs-to-reproduce-human-linguistic-variability-in-next-word-prediction
	TACL	Campos et al. (2024)	NA
	ACL Findings	Gomes et al. (2025)	https://github.com/gecgomes/Conformal_CLIPScore
	Preprint	Ilia and Aziz (2024b)	https://github.com/evgeniael/probar

Table 1: Research outputs (manuscripts and code) from T5.1

Period	Venue	Paper	Code
RP1	XAI4CV Workshop	Nalmpantis et al. (2023)	https://github.com/AngelosNal/Vision-DiffMask
	ACL	Treviso et al. (2023)	https://github.com/deep-spin/crest/
	ACL	Moghe et al. (2023)	NA
	ACL	Rei et al. (2023)	https://github.com/Unbabel/COMET/tree/explainable-metrics
	TACL	Guerreiro et al. (2024)	https://github.com/Unbabel/COMET/
RP2	Preprint	Wang et al. (2024c)	NA
	ACL	Wang et al. (2025)	https://github.com/weixuan-wang123/INCLINE
	Preprint	Eikema et al. (2025a)	NA
	AISTATS	Campos et al. (2025)	https://github.com/deep-spin/sparse-activations-cp
	COLING	Cilleruelo et al. (2025a)	https://github.com/ilyocoris/generics_are_puzzling
	EMNLP Findings	Treviso et al. (2024)	http://huggingface.co/sardinelab/xTower13B
	ACL	Zaranis et al. (2025)	NA
	ACL	Fucci et al. (2025)	https://github.com/hlt-mt/speech-translation-gender
	NAACL	Mire et al. (2025)	https://github.com/joel-mire/rm-dialect-biases

Table 2: Research outputs (manuscripts and code) from T5.2

Period	Venue	Paper	Code
RP1	EACL	Guerreiro et al. (2023d)	https://github.com/deep-spin/hallucinations-in-nmt
	ACL	Guerreiro et al. (2023b)	https://github.com/deep-spin/ot-hallucination-detection
	TACL	Guerreiro et al. (2023a)	https://github.com/deep-spin/lmt_hallucinations
	EMNLP	Farinhas et al. (2023)	https://github.com/deep-spin/translation-hypothesis-ensembling
	EAMT	Glushkova et al. (2023)	https://github.com/deep-spin/robust_MT_evaluation
RP2	EMNLP	Ranaldi et al. (2025)	No
	EMNLP	Ranaldi et al. (2024)	No
	ACL Findings	Zhu et al. (2024b)	https://github.com/NJUNLP/QAlign
	Preprint	Zhu et al. (2024a)	NA
	TACL	Peters and Martins (2025)	NA

Table 3: Research outputs (manuscripts and code) from T5.3

2 Task 5.1: Uncertainty-aware generation and conversational QE (UVA*, IT, UEDIN, UNB)

Proposal

The key highlights from the proposal are listed below.

5.1a uncertainty-awareness: UTTER's systems should be "aware" of their own limitations (*e.g.*, in order to adequately handle ambiguous or out-of-domain inputs).

5.1b interpretable uncertainty: for smooth interactions with system users, we will develop and test methods to express uncertainty in a human-readable and verifiable fashion.

Summary of completed work

The original plan was to work on this task from the beginning of the project till month 24, but, given the increased relevance of uncertainty-awareness and conversational tools to contemporary literature and efforts, we have extended this task to month 36. We document further progress along the two dimensions above (*i.e.*, uncertainty-awareness and interpretable uncertainty).

As in RP1, our contributions to this task are organised under 4 sub-themes: statistical evaluation of text generators (5.1a), improved text generation (5.1a), assessing and editing the parametric knowledge of language models (5.1a), and interpretable uncertainty quantification (5.1b). We contributed methodology, data, software and empirical observations that advance the state-of-the-art.

2.1 Statistical Evaluation of Text Generators

Natural language generators are built upon probabilistic models which inherently pack a highly structured representation of uncertainty about responses given a prompt. Our work in this dimension of T5.1a was developed in RP1. We contributed to theoretical understanding and to the practical evaluation of text generators in settings of high human production variability (e.g., due to input ambiguity or under-specification), as it is commonly the case in NLG. The following three outputs were summarised in D5.1, together they've gathered nearly 100 citations, with the last one responsible for about half of them:

- Interpreting Predictive Probabilities: Model Confidence or Human Label Variation? (Baan et al., 2024, see D5.1, Section 2.1.1).
- Predict the Next Word: ⟨*Humans exhibit uncertainty in this task and language models* _____⟩ (Ilia and Aziz, 2024a, see D5.1, Section 2.1.2).
- What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability (Giulianelli et al., 2023, see D5.1, Section 2.1.3).

2.2 Improved Text Generation

Text generation is performed by combining a probabilistic model of responses given a prompt, and a decision rule (or decoder), that is, an algorithm that explores the probability distribution and elects an output to show to the user. In RP1, we contributed to a better understanding of failure modes of main stream text generation algorithms and introduced novel algorithms for machine translation. The following two outputs were summarised in D5.1; to date, they have gathered about 20 citations (mostly due to the second output):

- The Effect of Generalisation on the Inadequacy of the Mode (Eikema, 2024, see D5.1, Section 2.2.1).
- Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models (Waldendorf et al., 2024, see D5.1, Section 2.2.2).

We continued to work in this dimension of T5.1a throughout RP2 and contributed novel decoding algorithms aiming at a) supporting open-ended text generation applications (§2.2.1), b) obtaining diverse responses (§2.2.2 and §2.2.3), and c) learning to better predict token-level variability (§2.2.4).

2.2.1 Structure-Conditional Minimum Bayes Risk Decoding

Minimum Bayes risk (MBR) decoding has emerged as a robust alternative to maximum likelihood–based generation strategies, showing consistent gains performance in neural machine translation. However, in open-ended tasks such as dialogue and instruction following, outcome spaces of generations may contain considerably more *latent structure*, potentially expressed in multiple clusters of similar outcomes. Using similarity-based utility functions, as is standard in machine translation, may result in the decoding algorithm compromising between clusters, potentially leading to suboptimal generations.

Data and models. We construct three datasets of outcome spaces with controlled structural variation, each covering a representative type of latent structure: dialogue acts, emotions, and response structure (i.e. brief responses, full paragraphs, lists and tables). For each context, we generate 25 responses per structure category, yielding 3,000 outcome spaces with a total of 350k candidate generations. For evaluation on real-world tasks, we additionally use AlpacaEval (Dubois et al., 2024) and MT-Bench (Bai et al., 2024), with unbiased candidate samples generated from OLMo2 13B (Walsh et al., 2025). Experimental code and the structural variation dataset are released at https://github.com/roxot/structure-conditional-mbr.

Methodology. To increase the sensitivity of MBR to structural variation, in Eikema et al. (2025b) we introduce *structure-conditional MBR*, a set of lightweight adaptations to MBR utilities designed to increase sensitivity to structural variation in the outcomes. We propose two new evaluation metrics, *cluster optimality* and *cluster-optimal rank correlation*, to assess whether MBR solutions respect latent structure, using our structural variation dataset as evaluation setup.

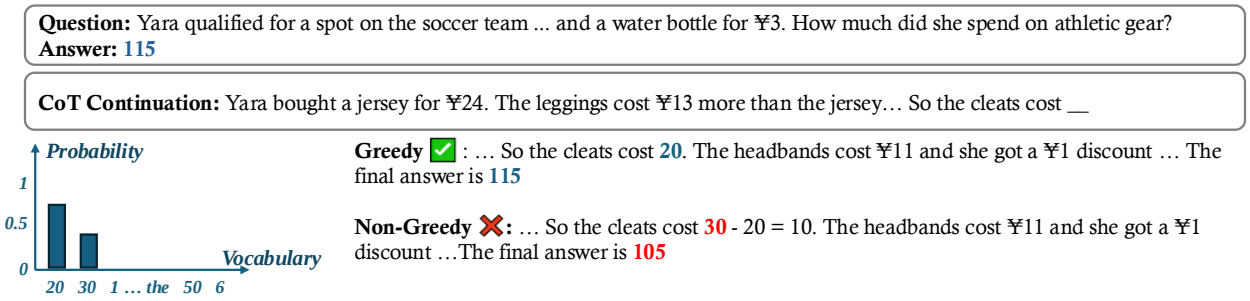


Figure 1: High sampling risk example for the current decoding position. **CoT Continuation** is generated by greedy decoding. **Greedy** continuation 20 results in a correct answer while **non-greedy** continuation 30 leads to an incorrect answer.

Findings. Our analysis reveals that standard utilities such as BLEURT and BERTScore achieve cluster-optimality in fewer than half of the cases, highlighting the need for structural adaptation. We propose three simple methods to adapt utility functions and show substantially improvements in structural sensitivity, with gains of over 30 percentage points in cluster-optimality. Applied to real-world benchmarks (AlpacaEval, MT-Bench), these approaches yield improvements of up to 13.7 percentage points in win rate against GPT-4o, demonstrating that structure-aware decoding improves both theoretical optimality and practical generation quality.

This work is reported in Eikema et al. (2025b).

2.2.2 Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs

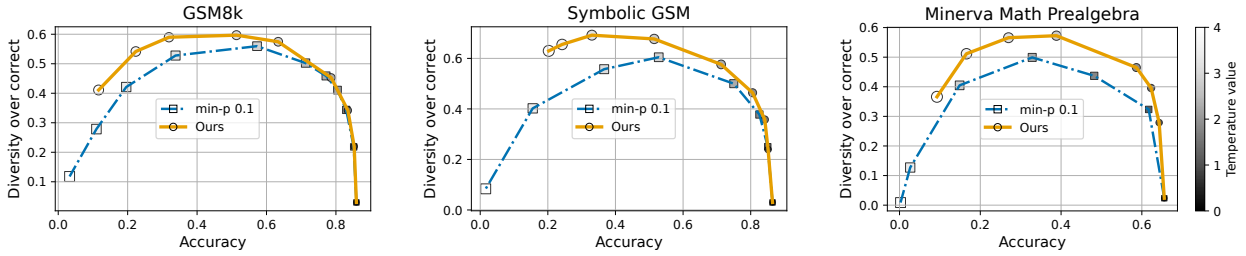
Temperature-based sampling is a common strategy for increasing diversity of large language model outputs. We demonstrate that, for tasks that require high precision, *e.g.*, mathematical reasoning, uncontrolled high temperature sampling, *e.g.*, min-p or top-p, degrades reasoning quality. We build a controlled approach that avoids this problem, leading to a better diversity/quality trade-off.

While arithmetic tasks are not an explicit focus of UTTER, they appear as instances of questions and answers in interactive assistant scenarios, *e.g.*, the *how many* questions in ELITR-Bench.

Data and models. Following common practice in the area, we evaluate on commonly used mathematical reasoning benchmarks:

- **GSM8K** (Cobbe et al., 2021): A grade school math problem-solving benchmark.
- **GSM-*Symbolic*** (Mirzadeh et al., 2025): An extended variant of GSM8K with symbolic templates, designed to provide a more reliable assessment of reasoning ability.
- **Minerva MATH** (Hendrycks et al., 2021): A dataset of competition-level mathematical problems. We conduct experiments on the PreAlgebra subset.

Methodology. Motivated by the concept of *regret* in reinforcement learning, we define **sampling risk** as a measure of the price paid in accuracy if choosing to sample (explore) instead of to decode



((a)) GSM8k **((b))** Symbolic GSM **((c))** Minerva (Prealgebra)

Figure 2: Selective sampling with our classifier improves the diversity-quality trade-off compared to the strong min$-p$ truncation baseline. On the x-axis, we report the accuracy, and on x-axis, we report the Diversity over correct samples. Size and color of the circles mark the temperature parameter.

the next greedy token (exploit) in a given prefix context x:

$$\text{s-risk}(x) := R(x) - \mathbb{E}_{v\sim p}\left[R\left([x, v]\right)\right], \tag{1}$$

where $[x, v]$ denotes the concatenation of the current prefix with a sampled next token, and $R(x)$ is the reward obtained by continuing from x with only greedy tokens until the stopping criteria is met. (For arithmetic tasks, R is a binary evaluation of the final answer.)

We then train a classifier to run alongside the LLM decoding and to estimate the sampling risk at the current context. At deployment, we use the output of this classifier to decide whether a context is safe for exploration, or if greedy exploitation is preferable.

Findings. As shown in Figure 2, our selective sampling approach outperforms the state-of-the-art top-p sampling strategy in terms of the diversity-quality trade-off: quality remains high when diversity is not needed, and degrades substantially less when temperature is increased.

We believe that this pilot study can impact the researrch and practice of controllable language generation for trustworthy assistants, since sampling accurate decoding paths is necessary in order to get sufficient candidates from which to identify ones that conform to the generation constraints.

This work is reported in Troshin et al. (2025a).

2.2.3 Asking a Language Model for Diverse Responses

Large language models increasingly rely on explicit reasoning chains and can produce multiple plausible responses for a given context. We study the candidate sampler that produces the set of plausible responses contrasting the ancestral (parallel) sampling against two alternatives: enumeration, which asks the model to produce n candidates in one pass, and iterative sampling, which proposes candidates sequentially while conditioning on the currently generated response set.

Data and models. Output diversity can lead to a deterioration in quality, but quality is generally hard to quantify. We therefore focus on simple math problem solving on GSM8K (Cobbe et al., 2021) math problems, where there is a unique correct answer that can be exactly verified. We evaluate the the Qwen3 family of models (Yang et al., 2025), chosen for their high reasoning performance, diverse range of model sizes. In our preliminary investigation, we observe that Qwen3

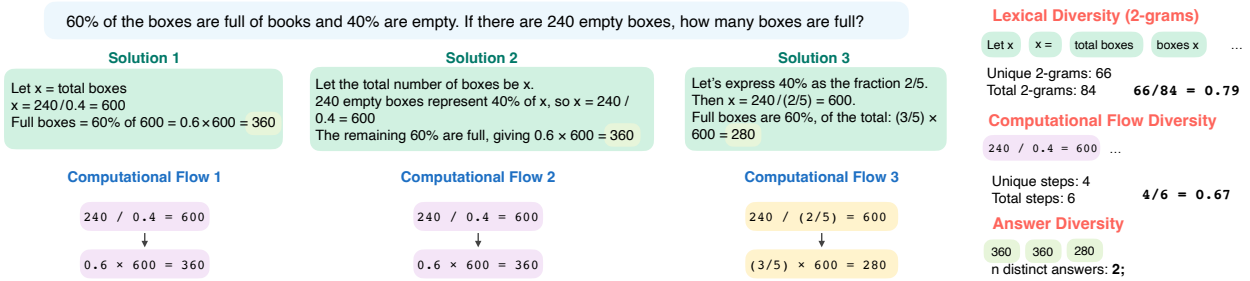


Figure 3: Example of a math problem with three responses, their computation flows, and the resulting metrics: lexical, computational flow and answer diversity.

models are able to follow our zero-shot instructions, and they show high accuracy in following the required output format. For our experiments, we use `Qwen3-{4B,8B,14B}` models with thinking generation mode on; and we use `Qwen3-4B-{Instruct/Thinking}-2507` released solely for non-thinking/thinking use-cases.

Methodology. In contrast to more conventional studies on diversity in generation, Ippolito et al. (2019), in this work, we take a substantially different approach and ask whether we can use the standard LLM generation pipelines to enable efficient non-independent sampling, by processing multiple candidates at the same time.

In particular, we are interested in a candidate sampler that:

(i) produces high-quality samples;

(ii) promotes response diversity;

(iii) scales efficiently as the number of responses increases;

(iv) is simple to use and relies on standard LLM decoding primitives.

We compare the commonly used **parallel** sampling strategy (ancestral sampling) with two alternative sampling strategies, which we define as **enumeration** and **iterative** approaches, and study them from the perspective of quality, diversity, and efficiency. Under matched budgets, we compare parallel, enumeration, and iteration samplers on quality, lexical and computation flow diversity, and efficiency (see Figure 3).

Findings. Our empirical results demonstrate that enumeration and iterative strategies result in higher diversity at comparable quality. Our findings highlight the potential of simple non-independent sampling strategies to improve response diversity without sacrificing generation quality.

This work is described in (Troshin et al., 2025b).

2.2.4 Learning to vary: Teaching LMs to reproduce human linguistic variability in next-word prediction

Natural language generation (NLG) tasks are often subject to inherent variability; e.g. predicting the next word given a context has multiple valid responses. While having language models (LMs)

that are aligned pluralistically, so that they are able to reproduce well the inherent diversity in perspectives of an entire population of interest is clearly beneficial, previous work shows that LMs do not reproduce this type of linguistic variability well. They speculate this inability might stem from the lack of consistent training of LMs with data reflecting this type of inherent variability. As such, we investigate whether training LMs on multiple plausible word continuations per context can improve their ability to reproduce human linguistic variability for next-word prediction.

Data and models. In our experiments, we use pre-trained GPT-2 (124M; Radford et al. (2019)) and instruction-tuned Mistral-7B-Instruct-v0.3 (7.25B; Jiang et al. (2023)) , which we refer to as Mistral-7B-IT. Both models are fine-tuned using Provo Corpus (Luke and Christianson, 2018), which contains 55 text passages (2687 total contexts). Each prefix is annotated with an average of 40 human annotations predicting the word following it.

Methodology. We employ techniques to fine-tune pre-trained LMs and instruction-tuned LMs. For the former, we alter the training signal, and for the latter we exploit a training data augmentation method to ensure that variability is observed. We employ these fine-tuning techniques for GPT-2 (Radford et al., 2019), a pre-trained model, and Mistral-7B-IT (Jiang et al., 2023), an instruction-tuned model.

Findings. When evaluating, by measuring divergence among empirically estimated human and model next-word distributions across contexts, before and after fine-tuning, we find that fine-tuning with multiple labels per instance improves those LMs' ability to reproduce linguistic variability, across contexts of varying open-endedness. With additional analysis and ablations, we measure performance when varying the number of training labels per instance, and we compare to models trained using majority labels. Moreover, with a preliminary analysis we measure the trade-off in performance in tasks that admit no plausible variability. For that, we handcraft a small evaluation dataset using a knowledge-based question answering dataset.

This work is reported in Groot et al. (2025).

2.3 Assessing and Editing What Models 'Know'

Because LLMs are unlike a typical data base, if they store any facts observed during pretraining, these ought to be stored in the LLM's parametric memory. Storage and retrieval of these facts are trainable parametric mechanisms that remain mostly opaque to practitioners and researchers alike. Our work in this dimension of T5.1a was developed in RP1. We contributed methods for assessing and maintaining the parametric knowledge of LMs. The following two outputs were summarised in D5.1; to date, they have gathered over 50 citations (mostly due to the second output):

- Assessing the Reliability of Large Language Model Knowledge (Wang et al., 2024b, see D5.1, Section 2.3.1).[1]
- Retrieval-augmented Multilingual Knowledge Editing (Wang et al., 2024a, see D5.1, Section 2.3.2).[2]

[1] First appeared in RP1, in preprint (Wang et al., 2023b).
[2] First appeared in RP1, in preprint (Wang et al., 2023a).

2.4 Interpretable Uncertainty Quantification

Uncertainty is typically represented by a probability distribution, with probability functioning as a mechanism to order events from most to least uncertain. Probability is, however, not always easy for humans to interpret, and this is also true for other summaries of uncertainty based on probability (*e.g.*, entropy). In RP1, we contributed towards more human interpretable forms of uncertainty quantification by disentangling uncertainty representations along aleatoric and epistemic dimensions, and by creating so-called (conformal) prediction sets. These contributions advance sub-goal 5.1b (interpretable uncertainty). The following three outputs were summarised in D5.1; to date, they have gathered more than 50 citations:

- Disentangling Uncertainty in Machine Translation Evaluation (Zerva et al., 2022, see D5.1, Section 2.4.1).
- Non-Exchangeable Conformal Language Generation with Nearest Neighbours (Ulmer et al., 2024, see D5.1, Section 2.4.2).
- Non-Exchangeable Conformal Risk Control (Farinhas et al., 2024, see D5.1, Section 2.4.3).

We continued to work in this dimension of T5.1 throughout RP2, we continued to contribute to the theory and applications of conformal prediction in NLP (§2.4.1 and §2.4.2), and also continued to contribute techniques towards more human interpretable forms of uncertainty quantification (§2.4.3).

2.4.1 Conformal Prediction for Natural Language Processing: A Survey

Acknowledging the rising importance of calibration of language model outputs and uncertainty metrics, we review the growing body of work on conformal prediction (CP) and its diverse applications in natural language processing. We describe the main families of approaches, including split conformal prediction, Mondrian CP, cross-conformal prediction, Venn–Abers predictors, and conformal risk control (CRC). We note that language modelling is a challenging task to apply CP to, due to the sequential model of language generation, which breaks the exchangeability assumption that underpins the statistical guarantees of CP and CRC. We thus further discuss extensions that relax exchangeability assumptions, enable conditional coverage, or improve uncertainty calibration, and highlight how these methods can be integrated into NLP pipelines to produce statistically valid prediction sets without requiring retraining of the underlying models.

We then examine **applications of CP** across several areas of NLP. For text classification and sequence tagging, CP has been shown to provide calibrated prediction sets and robustness in both binary and multilabel tasks. In natural language generation tasks such as machine translation and summarization, CP supports prediction sets that capture multiple valid outputs and help obtain better representations of uncertainty. CP has also been used to improve efficiency by enabling safe early exiting during decoding, pruning intermediate outputs while preserving coverage guarantees. Finally, we review work that leverages CP for uncertainty-aware evaluation, where calibrated prediction intervals provide a more principled assessment of model reliability.

We identify several open challenges for CP in NLP. These include designing prediction sets that are usable and informative in interactive systems, distinguishing between model uncertainty and genuine human label variation in tasks with multiple correct answers, and ensuring fairness by

maintaining equal coverage across population groups. We also point to the difficulty of calibrating CP methods with limited or imbalanced data, and to the need for deeper integration of CP into uncertainty-aware evaluation frameworks. Together, these observations establish CP as a promising and still developing methodology for trustworthy and user-aligned NLP applications and language models.

This work is reported in Campos et al. (2024).

2.4.2 A Conformal Risk Control Framework for Granular Word Assessment and Uncertainty Calibration of CLIPScore Quality Estimates

Beyond quantifying uncertainty, it is key to develop evaluation methods that are not only accurate but also trustworthy, particularly by quantifying uncertainty in ways that can guide user decisions. We explore this direction, focusing on vision-language models and the image-captioning use-case, which have seen rapid developments lately. Existing image captioning evaluation metrics, such as CLIPScore, provide global quality estimates but lack the ability to identify specific word-level errors and to produce calibrated confidence intervals. This limits their interpretability and reliability in downstream applications where users need fine-grained feedback and well-calibrated uncertainty estimates.

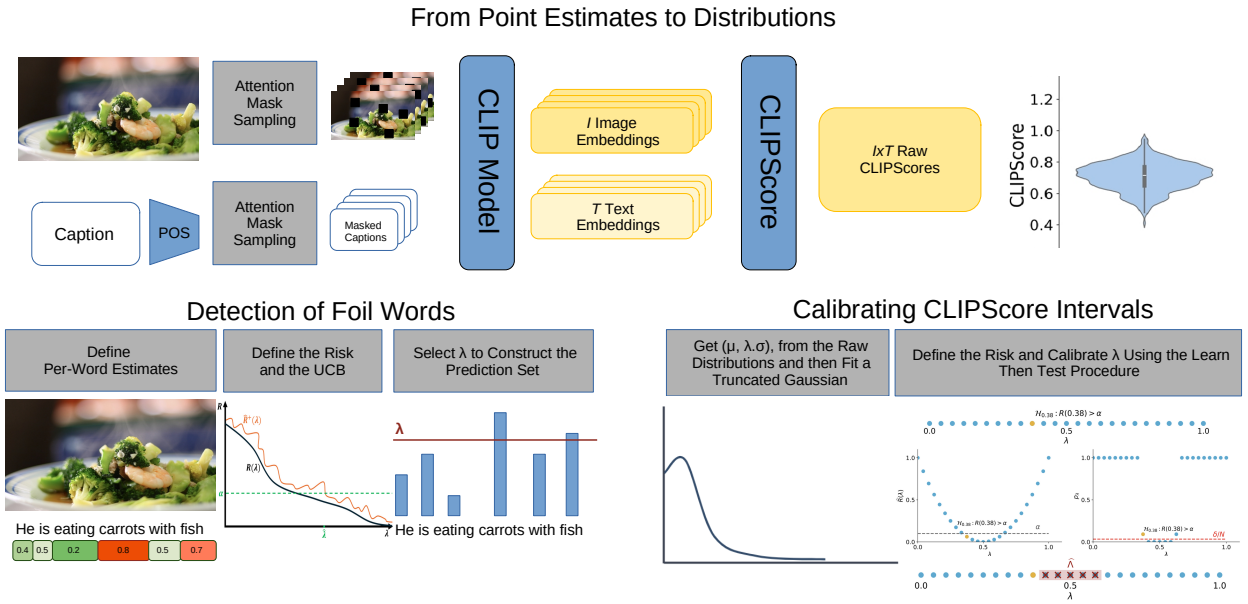


Figure 4: General overview of the proposed approach, using conformal risk control over CLIPScore values for two particular applications, namely the detection of foil words and the production of calibrated CLIPScore intervals.

Method. To address these limitations, we proposed a conformal risk control framework that augments CLIPScore with granular word-level assessment and uncertainty calibration, as shown in Figure 4. Our approach is model-agnostic and relies on stochastic attention mask sampling over CLIP encoders to generate score distributions and thus provide a flexible proxy to uncertainty estimation. To detect potentially erroneous words, we employ the same idea of stochastic mask sampling, to generate a score for each word, representing the score variance for each word,

based on the underlying assumption that quality scores will diverge more when masking erroneous words. To calibrate the detection of such words with respect to different performance metrics (e.g., false positive rate), we integrate these word-level scores into a conformal risk control framework, to determine "foil" word thresholds that satisfy user-decided performance levels. We the proceed to employ conformal risk control to calibrate the confidence scores estimated over full captions, aiming to provide statistical guarantees on a risk function tailored to evaluation, namely the **Uncertainty Pearson Score** (UPS; Glushkova et al. (2021)), which measures the correlation between prediction errors and the estimated uncertainties.

Findings. We evaluated our method on FOIL-it (Shekhar et al., 2017), FOIL-nocaps (Petryk et al., 2024), and Rich-HF (Liang et al., 2024), covering both multi-class and multi-label error detection. Across datasets, we found that our framework detects foil words effectively, improves the correlation between uncertainty estimates and prediction errors, and enhances the reliability of CLIPScore as a quality metric. These findings demonstrate that conformal prediction can provide a principled path towards uncertainty-aware and interpretable evaluation, directly advancing UTTER's goal of building user-aligned and reliable assessment tools for multimodal generation.

This work is reported in Gomes et al. (2025).

2.4.3 Revisiting the role of Variability in Uncertainty Quantification

With the broader use of language models (LMs) comes the need to estimate their ability to respond reliably to prompts. Uncertainty quantifiers (notions of confidence and entropy, *i.a.*) can be used to reject generated responses that are likely to be incorrect (*i.e.*, selective prediction). The widely used semantic entropy (Kuhn et al., 2022) regards *semantic* variation amongst sampled responses as an indicator that an LM is likely to err. We argue that semantic homogeneity need not imply correctness, whereas semantic variability need not imply error—with the latter being especially intuitive in open-ended settings, where prompts elicit multiple *adequate* but semantically distinct responses (Giulianelli et al., 2023). To accommodate this more general setting, rather than judging a model's reliability by its confusion among semantically distinct responses, we propose to annotate sampled responses for their adequacy to the prompt, as judged by an external reward model. We then estimate the Probability the model assigns to Adequate Responses (ProbAR) and regard that as an indicator of the model's reliability when responding to a given prompt; see Figure 5. We show ProbAR's potential as a viable alternative to variation-based quantifiers by evaluating (manually and automatically) ProbAR, implemented using Mistral models, in selective prediction, for QA and next word prediction.

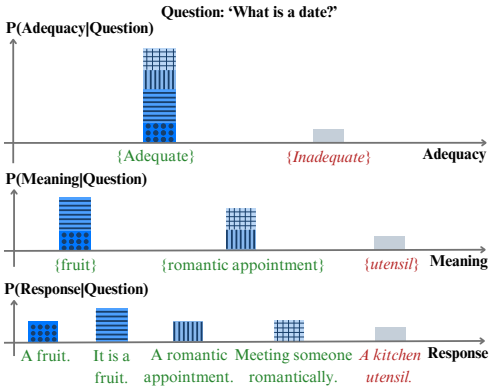


Figure 5: Bottom: samples drawn from an LLM, given a question; the model exhibits high entropy over surface forms, but some responses are semantically equivalent. Middle: responses are clustered by meaning); while this representation still exhibits high *semantic entropy*, probability concentrates on answers wrt different but plausible interpretations of the question. Top: responses are grouped as a function of their adequacy to the prompt, as we propose; now it is clear that, however semantically diverse, responses are mostly adequate to the prompt—we regard the probability of adequate responses as an expression of the model's reliability.

Data and models. We test ProbAR using Abg-COQA (Guo et al., 2021), a reading comprehension QA (RCQA) dataset with ambiguous and unambiguous questions: 994 context-ambiguous question pairs (741, 130, 123 in training, development and test sets, resp.) and 1894 non-ambiguous prompts (962 and 932 in development and test sets, resp.). In the paper, we also report experiments on AmbigQA (Min et al., 2020), which is a knowledge based QA (KBQA) dataset and on Provo Corpus (Luke and Christianson, 2018), a dataset for next-word prediction in English. We generate responses from OPT (2.7b, 6.7b, 13b and 30b; Zhang et al. (2022)) and Mistral instruction-tuned models (`7B-Instruct-v0.2`, `Nemo-Instruct-2407`, `Small-Instruct-2409`; referred to as Mistral 7b, 12b and 22b resp.; Mistral (2024)). For each prompt, we obtain 10 unbiased samples for uncertainty quantification.

Method. For each task, ProbAR requires an adequacy classifier, which we designed by prompting Mistral12b (or Mistral22b and Mistral12b, in different experiments). In all cases, we prompt an LLM to perform adequacy judgments; in experiments we compare that to an upperbound where these judgements are performed by a human annotator, for a small subset of the Abg-COQA prompts. For Abg-COQA, we have an LM perform adequacy judgements, with the passage and a question-answer pair provided as context and the LM prompted to generate true if the answer to the question is plausible given the passage, or false if not. For AmbigQA, we prompt an LM to generate true if a response is adequate given the question with respect to the LM's own training data. For NWP, we prompt an LM to generate true if a response is plausible given the context and false otherwise. We experiment with using Mistral12b, Mistral22b and Mistral12b for adequacy classification in Abg-COQA, AmbigQA and Provo resp., and also experimented with using the same LLM generator as a classifier (prompting it to perform adequacy classification, that is). As baselines for comparison, we employ Shannon entropy (E), estimated via MC, semantic entropy (SE), including their versions normalised to account for the size of the observed outcome space for each prompt, and a variant of P(True) (Kadavath et al., 2022) which estimates the confidence of a response (*i.e.* greedy) by prompting for adequacy (instead of correctness) and assessing the associated token's probability (P(Adequate). We evaluate the models on AUROC: for the QA tasks, we obtain the greedy decoding and regard that as the LM's prediction, as typically done for QA (Kuhn et al., 2022). For NWP, we choose (at random) one of the sampled responses and regard that as the LM's prediction. AUROC requires a notion of correctness by which to criticise the LM's prediction; for QA tasks, we follow Lin et al. (2024) and let `gpt3.5-turbo` (via the OpenAI API) determine if the LM prediction is correct against the multiple references from the QA datasets; for Provo Corpus, we regard the LM prediction as correct if it exactly matches one of the prefix's reference answers.[3]

Findings. In Figure 6, we concentrate on the subset of AbcCOQA for which we obtained manual labels (for semantic clusters in SE, adequacy judgements in ProbAR, and the correctness of greedy responses in the evaluation protocol). In all plots, we observe AUROC values for different UQs across OPT model sizes, with SE's and ProbAR's internal decisions replaced by human judgement. On the left, AUROCs are computed using hand-labelled assessments for the greedy responses' correctness, hence we observe results in a setting free of errors in the evaluation protocol. ProbAR (M) and Norm.SE(M) represent the 'upper-bounds' for SE and ProbAR, respectively, computed using

[3] For this project, UTTER (UVA) purchased OpenAI API credits for evaluation; the complete evaluation costs were under 50 EUR in credits. Training and evaluation performed by open models were done at UVA clusters, with no extra costs to the project.

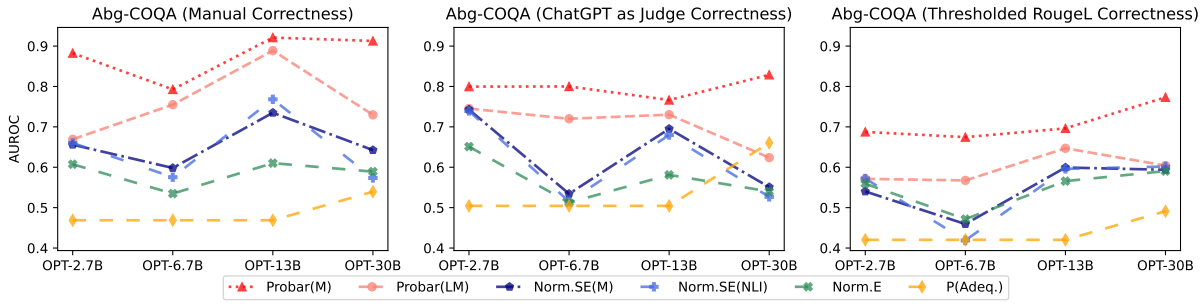


Figure 6: AUROC values for AbgCOQA's 50 manually annotated contexts. Left: AUROC is computed using manual correctness annotations for the greedy; middle and left: greedy's correctness was automated using `gpt3.5-turbo` and thresholded Rouge-L resp., where the latter criterion errs more often.

the manual adequacy and semantic equivalence annotations. ProbAR outperforms all baselines, both its upperbound ProbAR (M) and its practical implementation ProbAR (LM). Beyond the fact that ProbAR (M) surpassed Norm.SE(M) for all OPT models by a large margin, it is noteworthy that ProbAR (LM) surpassed Norm.SE(M). On the middle and right plots, we see the results for the same subset, but using `gpt3.5-turbo` and `RougeL` resp. to automate the evaluation protocol (*i.e.*, assess correctness of the greedy responses). We can see that errors in evaluation have a greater impact on more informative quantifiers (*i.e.*, with higher AUROC values), with the impact growing when the automated algorithm used has higher error rates (F1 for `gpt3.5-turbo` around 0.9 and for `RougeL` 0.8. Better quantifiers get negatively impacted by misclassifications of plausible decodings they accept with high confidence, while worse uncertainty quantifiers, which wrongly abstain from answering, dodge these errors. In the paper, we show in a large scale automated evaluation that these findings generalise across the three tested task/datasets.

To conclude, we demonstrated how semantic variation (or lack of) need not robustly predict propensity for error (or correctness), and how ProbAR better informs us whether the model can reliably respond to a prompt, regardless of the prompt's open-endedness.

This work is described in Ilia and Aziz (2024b).

3 Task T5.2: Explainability (UVA*, IT)

Proposal

The key highlights from the proposal are listed below.

5.2a explaining predictions: trustworthy language technology should provide correct attributions and explanations of their output (*e.g.*, meeting assistant should provide pointers into specific timestamps or quotes from the meeting to justify action items).

5.2b transparent evaluation: besides explaining predictions of a trained model, we will adapt models of quality estimation and machine translation evaluation making them more easily amenable to human interpretation.

Summary of completed work

In the original plan, this task was scheduled to begin from month 12 and extend until the end of the project. We decided to schedule it earlier, already from the beginning of the project, and it extended till the project's end. Our contributions are organised under 2 themes: explaining predictions (5.2a) and transparent evaluation (5.2b).

3.1 Explaining Predictions

Explaining predictions of a trained model is often done, amongst other things, in terms of attribution methods (such as input or data attribution), causal and counterfactual interventions, analysis of training dynamics and/or inference-time activations, natural language explanations, and sparsity. In RP1, we contributed to faithful input attribution in Transformers encoders and to explaining and improving text generation. The following two outputs were summarised in D5.1; to date, they have gathered about 30 citations (mostly due to the second output):

- VISION DIFFMASK: Interpretability of Computer Vision models with Differentiable Patch Masking (Nalmpantis et al., 2023, see D5.1, Section 3.1.1).
- A Joint Framework for Rationalization and Counterfactual Text Generation (Treviso et al., 2023, see D5.1, Section 3.1.2).

We continued to work in this dimension of T5.2 throughout RP2 and continued to contribute to explainability of Transformer models but now with a focus on multilingual LLMs (§3.1.1 and §3.1.2), verbalising an LM's confidence in a generated response (§3.1.3) and sparsity-driven interpretability in conformal prediction (§3.1.4).

3.1.1 Sharing matters: Analysing neurons across languages and tasks in LLMs

LLMs exhibit strong performance across languages and tasks but remain difficult to interpret, particularly in multilingual settings. Prior work has mainly examined neuron behavior in monolingual contexts, leaving open the question of how neurons are shared or specialized across languages and tasks. In Wang et al. (2024c), we address this gap by analyzing the extent to which neurons are activated similarly or differently when processing multilingual input, with the aim of improving model explainability.

Data and Models. We evaluate five open-source multilingual transformer models (BLOOMZ-7b, BLOOM-7b, LLaMA2-7b-chat, XGLM, and mT0) multilingually on natural language inference (XNLI), fact probing, and cross-lingual knowledge editing. Using translated datasets to ensure comparability across languages, the tasks are applied across ten languages: English (en), German (de), Spanish (es), French (fr), Portuguese (pt), Russian (ru), Thai (th), Turkish (tr), Vietnamese (vi), and Chinese (zh).

Methodology. Neuron activation patterns in the feed-forward layers are categorized into four groups: all-shared (activated across all languages for a given example), partial-shared, specific (language-unique), and non-activated. Two attribution metrics are introduced: the Contribution

Score, measuring overall impact on outputs, and the Effective Score, measuring influence on correct predictions. Ablation experiments further test the functional role of different neuron groups.

Findings. Results show that neuron sharing is highly task- and example-dependent, with few neurons maintaining the same category across all contexts. Sharing does not fully follow linguistic family relationships, and increasing the number of languages tends to raise partial sharing. Crucially, all-shared neurons are consistently the most influential, carrying the highest attribution scores and yielding up to a 91.6% decrease in performance when deactivated. These results underscore the central role of shared neurons in multilingual LLM explainability.

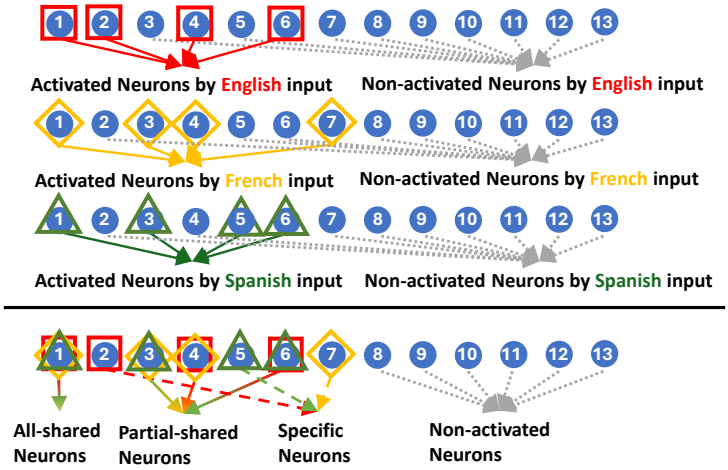


Figure 7: A comparison of neuron analysis with different type designs in multilingual settings with the same semantic input, in which we define four types of neurons in one layer of LLM.

This work is described in Wang et al. (2024c).

3.1.2 Bridging the Language Gaps in Large Language Models with Inference-Time Cross-Lingual Intervention

LLMs achieve state-of-the-art performance across many tasks, yet their abilities remain uneven across languages. High-resource languages such as English benefit from abundant training data, while low-resource languages lag behind significantly. Prior solutions, including additional pretraining or fine-tuning, often require substantial computational cost and data resources. In Wang et al. (2025), we address this challenge by proposing a more explainable and efficient approach: an inference-time mechanism that does not alter the base LLM parameters.

Methodology. Our method, Inference-Time Cross-Lingual Intervention (INCLINE), relies on alignment matrices learned between a low-resource and a high-resource language using parallel data. For each layer of the model, a transformation is trained via least-squares optimization to map hidden states of the source language into the target language's representational space. At inference time, these matrices are applied to the model's activations, effectively steering low-resource language representations toward the stronger processing pathways of high-resource languages. Importantly, this procedure requires no retraining or fine-tuning of the original model, making the intervention lightweight and transparent.

Findings. Experiments conducted across nine benchmarks and five different LLMs show that INCLINE consistently improves performance across tasks, with gains of 1.02 to 9.46 points over strong baselines at negligible computational cost. Further analysis demonstrates that higher-quality parallel data, such as human-translated sentences, enhances the effectiveness of alignment, with the INCLINE-FDEV variant outperforming the standard version. Overall, our approach highlights the value of directly manipulating internal representations at inference time, offering an interpretable and resource-efficient path to narrowing multilingual disparities.

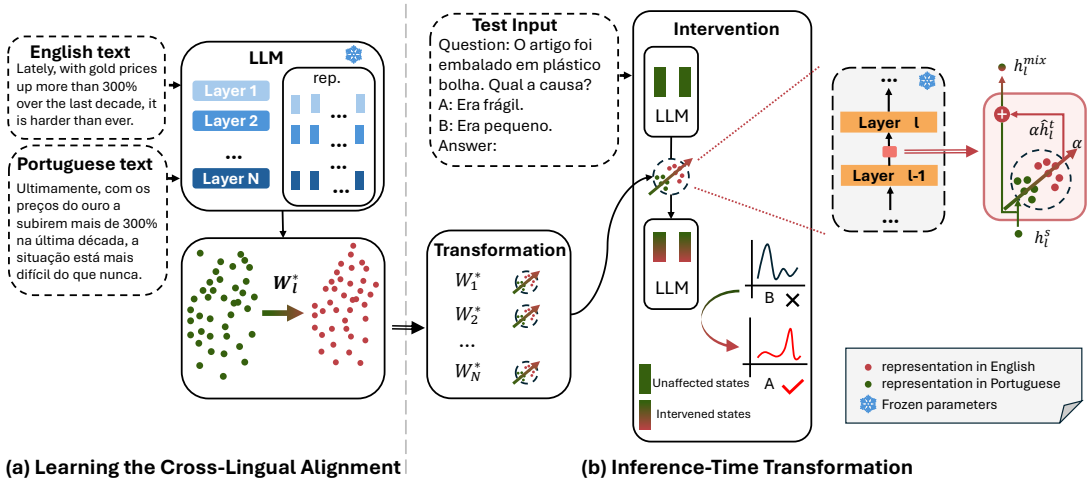


Figure 8: Our framework involves two steps: (a) Learning the Cross-Lingual Alignment: sentence representations from a parallel dataset are used to train alignment matrices that map source (Portuguese) representations to the target (English) representations. (b) Inference-Time Transformation: this step adapts the source representations from downstream tasks into the target representation space using the alignment matrices.

This work is described in Wang et al. (2025).

3.1.3 Teaching Language Models to Faithfully Express their Uncertainty

Large Language Models (LLMs) often fail to communicate when they are uncertain, creating a gap between their internal confidence and the decisiveness of their verbal output. This mismatch can lead to user over-reliance on incorrect answers. To address this, in our work we propose *Faithful Uncertainty Tuning (FUT)*, a fine-tuning method that trains LLMs to insert linguistic uncertainty markers consistent with their own intrinsic confidence, see Figure 9.

Data and models. We evaluate FUT on OLMo2 7B, OLMo2 13B, and Tulu3 Llama 8B (Walsh et al., 2025; Lambert et al., 2025) using question-answering (QA) datasets: PopQA (Mallen et al., 2023), Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). To supervise uncertainty expression, we construct training data with uncertainty hedges either *interwoven* into responses or *postfixed* at the end of the response.

Methodology. FUT fine-tunes models to align linguistic hedges with their intrinsic uncertainty, measured via sample consistency across generations. During training, hedges are injected into responses in a manner that reflects the variability of sampled answers. This allows the model to

verbalize uncertainty in a way that faithfully mirrors its internal distribution, without explicitly optimizing for correctness.[4]

Findings. FUT significantly improves model faithfulness, measured by the conditional mean faithful generation (cMFG) score (Yona et al., 2024), which quantifies whether expressed certainty matches sample-level variability. Our models improve from near-random faithfulness (cMFG $\approx$ 0.50) to state-of-the-art levels (cMFG $\approx$ 0.75–0.80). These gains generalize across QA datasets. Importantly, FUT does not substantially alter underlying QA performance or the distribution over semantically distinct answers. Finally, we also train models to predict numerical confidence directly, achieving similar performance and showing the applicability of FUT for both linguistic and numeric uncertainty communication.

This work is reported in Eikema et al. (2025a)

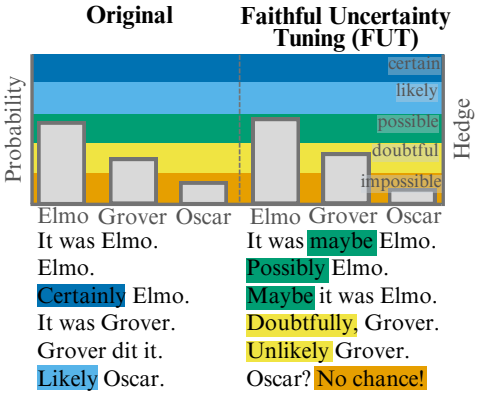


Figure 9: We want to know *Who pushed Big Bird?* and, for transparency's sake, we expect a good response to be suggestive of uncertainty in the model's state of knowledge. Each plot shows the beliefs of a model expressed as probabilities over clustered responses, with the actual responses shown below the plot. We segment the vertical axes in 5 intervals, and highlight hedge phrases that humans regard as coherent with those intervals of probabilistic belief. The original model (left) generates responses that are either unhedged or hedged unfaithfully (*e.g.*, the model's belief in *Elmo* is in the ballpark of 'possible', but responses blaming Elmo suggest 'certainty'), leading to misinformation. Faithful uncertainty tuning (FUT; right) adapts the model so that responses are faithfully hedged while closely preserving the original beliefs.

3.1.4 Sparse Activations as Conformal Predictors

Conformal prediction provides strong distribution-free guarantees, but its efficiency depends critically on the choice of non-conformity scores. At the same time, sparse activation functions such as sparsemax and γ-entmax can directly yield prediction sets through their supports, without requiring explicit thresholding, since they assign non-zero probabilities only to a subset of labels. We aimed to connect these two research directions, showing how sparse activations can naturally serve as conformal predictors, thus bridging uncertainty quantification and sparse probabilistic modeling.

Data and models. We relied on computer vision and text classification benchmarks. Specifically, we used CIFAR10, CIFAR100, and ImageNet datasets (Krizhevsky, 2009; Deng et al., 2009) for vision tasks, training or fine-tuning convolutional neural networks and vision transformers (ViT; Dosovitskiy et al. (2021)). For text classification, we employed the 20 Newsgroups dataset (Mitchell, 1999), fine-tuning a BERT-base model. We com-

[4] For this output, UTTER (UVA) purchase OpenAI API credits for certain aspects of the evaluation, which required closed API models for conformity with the best community standards. The costs were about 1000 EUR (a detailed account of costs can be found in the paper's appendix). Here's an approximate summary: there are 3 models tested in 4 settings, each evaluation costs on average 25 USD, and we designed 4 main experiments. Experiment 1: all models evaluated on PopQA $\approx$ 300 USD. Experiment 2: 1 model evaluated on NQ and TriviaQA $\approx$ 200 USD. Experiment 3: analysis of impact of 2 hedge schemes for 1 model in 1 dataset $\approx$ 50 USD. Experiment 4: analysis of impact of decoding strategy for 1 model, 1 dataset and 2 decoding strategies $\approx$ 50 USD. We also used some 250 USD for some evaluations in code development and tests of code and infrastructure, amounting to about 850 USD before taxes.

pared our proposed methods against standard conformal approaches such as InvProb and RAPS Angelopoulos et al. (2021).

Methodology. We propose a methodology to obtain novel non-conformity scores that link conformal prediction to temperature scaling of sparse activations. We established the equivalence between sparsemax supports and conformal prediction sets when temperature is calibrated using conformal procedures. We then generalized this result to the broader γ-entmax family, showing that by properly defining the non-conformity function, prediction sets correspond exactly to γ-entmax supports with calibrated temperature. This construction inherits conformal coverage guarantees, while introducing sparsity and interpretability.

Findings. Our experiments showed that the proposed γ-entmax conformal predictors achieve the required marginal coverage guarantees while being competitive in efficiency and adaptiveness. The opt-entmax variant, which tunes γ on calibration data, consistently yielded smaller prediction sets than baselines while maintaining coverage. Sparsemax tended to produce larger sets but still achieved valid coverage, whereas log-margin often improved adaptiveness and singleton prediction coverage. Overall, our results demonstrate that conformalizing sparse activations expands the family of non-conformity scores with practical, efficient, and interpretable alternatives, effectively connecting conformal prediction and temperature scaling.

This work is described in Campos et al. (2025).

3.2 Transparent Evaluation

Automatic evaluation protocols are powered by trainable metrics, often built upon blackbox components such as pretrained LLMs. In RP1, our contributions included a novel paradigm for fine-grained evaluation of MT, an analysis of blackbox neural MT evaluation metrics, and a state-of-the-art approach for evaluation via fine-grained error detection. The following three outputs were summarised in D5.1; to date, they have gathered nearly 200 citations (mostly due to the third output):

- Extrinsic Evaluation of Machine Translation Metrics (Moghe et al., 2023, see D5.1, Section 3.2.1).
- The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics (Rei et al., 2023, see D5.1, Section 3.2.2).
- xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection (Guerreiro et al., 2024, see D5.1, Section 3.2.3).[5]

We continued to work in this dimension of T5.2 throughout RP2 and continued to contribute to fine-grained evaluation (§3.2.1) including work on bias in MT quality estimation (§3.2.3), in machine translation (§3.2.4) and in reward modelling (§3.2.5), we also contributed a novel LLM tailored to transparent evaluation of translation (§3.2.2).

[5] First appeared in RP1, in preprint (Guerreiro et al., 2023c).

3.2.1 Generics are puzzling. Can language models find the missing piece?

In this paper (Cilleruelo et al., 2025a) apply LLMs to the study of *generics*. Generics are sentences that express generalisations without making use of explicit quantifiers. Examples of generics are *ravens are black* or *ticks carry lyme disease*. Generics are an interesting object of study in computational linguistics because they still lack an agreed account of their semantics. In many examples they can be paraphrased by quantified statements but the actual quantifier varies greatly, and generics admit exceptions.

In order to study generics, we first extract a small corpus of generics and quantified sentences used in context, which we call *congen*. We then introduce a new metric, *p-acceptability*, which uses LLM surprisal in order to select the most appropriate quantifier for a sentence (where a generic is included as the null quantifier). By analysing the predictions of p-acceptability against the actual quantifier, we show that it captures semantic intuitions.

We then introduce three experiments which analyse generics in context using LLMs. In the first experiment we look at which quantifier is most suitable to replace the generic, finding that this varies across sentences with *all* and *most* being the most common with about 40% each. To determine whether generics are context-sensitive we look at how the predictions of p-acceptability vary as we increase the context length. We find that it is easier to predict generic sentences when more context is provided, but that this is not the case for quantifiers. Finally we look at a particular type of generics that are used to express stereotypes (often negative) about specific groups. We show that *all* is the most likely replacement for a negative generic stereotype, as opposed to *most* for a postive example.

This work is described in (Cilleruelo et al., 2025a). The follow-up work described in (Cilleruelo et al., 2025b), which concerns dataset construction, is described in detail in D2.2.

3.2.2 xTower: A Multilingual LLM for Explaining and Correcting Translation Errors

We introduce **xTower**, a translation-oriented LLM built on top of TowerBase-13B (Alves et al., 2024) to **explain** span-marked translation errors and **propose** a corrected translation. The model is obtained by distilling GPT-4 explanations on MQM-annotated WMT22 data (EN→DE, EN→RU, ZH→EN), followed by multilingual finetuning mixed with TowerBlocks MT data. Prompts present an "annotated translation" (error spans with severities) and adopt an explanation-then-correction format, which xTower handles in referenceless or reference-based modes. Crucially, xTower is agnostic to the source of spans—human or automatic (e.g., xCOMET)—and can thus plug into existing QE pipelines.

We evaluate on WMT23 MQM test sets for EN→DE, HE→EN, and ZH→EN, measuring (i) *relatedness* of explanations to the marked spans and (ii) *helpfulness* for understanding the error and guiding a fix, via expert human annotation. Relatedness is higher when spans are human-labeled (about 4.3 on a 0–6 scale) than when predicted by xCOMET (about 3.2), confirming the impact of span quality. Helpfulness scores average 4.4–4.6 for error understanding and 3.3–3.9 for guidance, indicating that explanations are generally informative and often suggest the path to a better translation.

Conditioning on spans and the generated explanations, xTower refines the translation and is assessed with COMET (primary), BLEURT, and COMET-Kiwi. Across language pairs, xTower improves over the original MT by roughly +1 to +3 COMET in referenceless setups; a hybrid

strategy that keeps the original translation when COMET-Kiwi is high and otherwise adopts the xTower correction yields further gains (up to ∼+2 COMET on HE→EN). Improvements concentrate on lower-quality inputs (original COMET ≤ 80). Moreover, xTower fixes the majority of span-marked errors and compares favorably to strong LLM baselines when used for post-editing.

Overall, xTower shows that free-text, span-grounded explanations can be both **useful to humans** and **actionable for models**, leading to measurable MT quality gains in multilingual settings and enabling practical hybrid protocols for cost-effective post-editing.

This work is presented in more detail in (Treviso et al., 2024).

3.2.3 Watching the Watchers: Exposing Gender Disparities in Machine Translation Quality Estimation

Quality estimation (QE) has become a cornerstone of modern machine translation (MT), enabling systems to operate without costly human references and powering MT-related tasks such as data filtering, reranking, and quality-aware decoding. However, as QE scores increasingly guide the development and deployment of MT models, less is known about whether these metrics introduce systematic biases that could propagate downstream. In particular, gender bias poses a serious concern, since many target languages encode grammatical gender, and translation choices may inadvertently privilege certain forms.

In this work, **we investigate gender bias in QE metrics** by testing 11 state-of-the-art models across multiple setups. The study covers widely used neural metrics—CometKiwi 22, CometKiwi 23 (XL/XXL), xCOMET (XL/XXL), and MetricX 23 (L/XL), as well as prompted LLM-based QE with GEMBA, using Mistral 7B, Gemma 2 9B, Llama 3.1 70B, and GPT-4o. To evaluate them, we rely on minimal-edit contrastive datasets designed for gender bias analysis: MT-GenEval, GATE, and mGeNTE. These corpora span eight target languages (e.g., Arabic, German, Spanish, Italian, Russian) and cover both binary contrasts (masculine vs. feminine forms) and gender-neutral translation alternatives. Our experiments further study intra-sentential and contextual cases, including both human-written references and machine-generated outputs.

Our findings reveal consistent and concerning patterns. Across datasets and languages, most **QE metrics systematically assign higher scores to masculine over feminine forms**, even when context clearly disambiguates gender. Similarly, gendered translations are generally preferred over neutral alternatives, suggesting that QE may penalize inclusive language. These biases are not only intrinsic to evaluation but also carry downstream consequences: they can skew data filtering pipelines and lead to MT systems that reinforce gendered preferences. While the extent of bias varies across models, no metric proved free of it.

Overall, this study highlights the importance of adapting quality estimation models to gender-aware phenomena to ensure fairness and representativeness.

This work is presented in more detail in Zaranis et al. (2025).

3.2.4 Different Speech Translation Models Encode and Translate Speaker Gender Differently

Speech translation (ST) is an established technology for breaking down language barriers, but concerns remain about how these systems handle speaker-related attributes such as gender. Since

many target languages require grammatical gender marking, ST systems must often infer gender from the speaker's voice and propagate it into translations. Failing to do so risks systematic biases, most notably a default toward masculine forms. Understanding how different ST architectures internally encode and use speaker gender is thus essential for ensuring fairness and inclusivity.

In this work, we **analyze gender encoding across three representative ST models**: a traditional encoder–decoder system (Transformer-based), and two newer speech+MT architectures, SeamlessM4T and ZeroSwot. Using probing methods—including a novel attention-based probe—we evaluate how well hidden states capture gender information. Probes are trained on MuST-C data annotated with self-declared pronouns, and tested both on generic utterances (MuST-C) and on gender-sensitive sentences from MuST-SHE, which require correct inflection of gender-marked terms in English→Spanish, French, and Italian. Translation quality is further assessed with COMET, while gender translation accuracy is measured with the MuST-SHE evaluation protocol.

Our results reveal striking differences across architectures. Encoder–decoder systems encode gender information robustly and achieve high accuracy in translating gender-marked terms (avg. 85.6%). By contrast, speech+MT systems encode little to no gender information, especially after the adapter layer, and default to masculine forms—achieving much lower feminine accuracy (14–51%). Interestingly, overall translation quality does not predict gender fidelity: the speech+MT models scored higher on COMET but worse on gender translation. Instead, we find a strong correlation between gender encoding and translation accuracy ($R^2 = 0.99$), showing that preserving acoustic cues is critical for fairer gender handling.

Overall, this study highlights a trade-off in modern ST architectures. While adapters enable flexible integration of speech and MT components, they tend to erase gender-related information, leading to systematic masculine defaults. We suggest that future research focus on adapter design and training strategies to preserve relevant speaker attributes, and on targeted evaluation protocols to mitigate gender bias in ST outputs.

This work is presented in more detail in Fucci et al. (2025).

3.2.5 Rejected Dialects: Biases Against African American Language in Reward Models

In this work, we investigated how reward models (RMs), which are central to preference alignment in large language models (LLMs), encode dialect and sociolect-related biases. Our motivation stems from the observation that preference datasets are often sourced from annotators non-representative of diverse speech communities, raising concerns about fairness, especially for marginalized dialects such as African American Language (AAL). There has been little to no research so far on biases introduced at the reward modeling stage, thus we designed a framework to quantify biases in RMs against AAL relative to White Mainstream English (WME).

We built our evaluation on two types of datasets. First, we used RewardBench (Lambert et al., 2024), a large-scale benchmark of preference datasets aligned with WME, and produced machine-translated AAL variants via VALUE (Ziems et al., 2022) and PhonATe (Deas et al., 2024). This yielded RB-AAL, which preserved semantic content while adding morphosyntactic and phonological features of AAL. Second, we incorporated the DeasGroenwold dataset (Groenwold et al., 2020; Deas et al., 2023), comprising human-authored AAL texts paired with human-translated WME equivalents. Together, these corpora enabled controlled comparisons of RM behavior across dialects.

We evaluated 17 state-of-the-art reward models, spanning RLHF-trained sequence classifiers (Christiano et al., 2017; Ouyang et al., 2022) and DPO-style direct preference models (Rafailov et al., 2024), to address three research questions:

1. Are RMs worse at predicting preferences in AAL vs. WME?
2. Do RMs systematically prefer WME over AAL completions?
3. Do RMs mirror the input dialect or steer toward WME?

Accuracy, effect size, and correlation analyses were applied to measure disparities, supported by statistical tests such as McNemar's test with Holm correction.

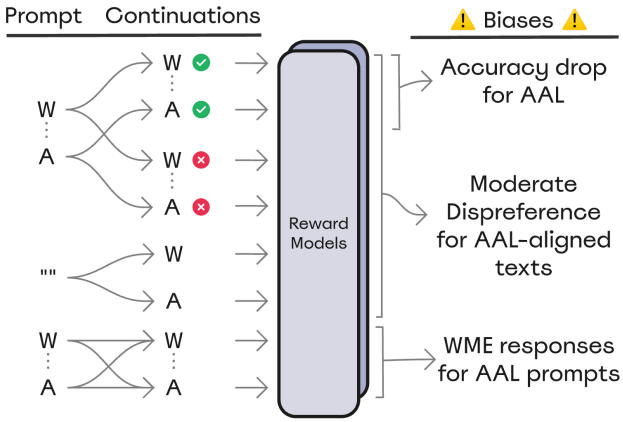


Figure 10: We analyze reward model scores for White Mainstream English (W) and African American Language (A) texts across various prompt-continuation settings. Vertical dotted lines indicate machine translations, and checkmarks/Xs indicate human preferences between alternatives. Our findings point to representational and quality-of-service harms for AAL speakers.

Our results revealed systematic biases as shown in Figure 10. RMs exhibited a statistically significant accuracy drop (–4% on average) when processing AAL texts, undermining their alignment with human preferences. In addition, most models consistently dispreferred AAL completions, with large positive effect sizes indicating bias toward WME. We also found that RMs overwhelmingly steered conversations toward WME, refusing to mirror AAL prompts. Case studies highlighted practical harms, including failures in refusal behavior and safety-related tasks under AAL input. These findings demonstrate that representational harms in reward models can cascade into quality-of-service harms, perpetuating raciolinguistic hierarchies in LLMs.

Overall, our study reveals critical limitations in preference alignment pipelines, showing that reward models encode structural biases against AAL. It underscores the importance of investigating fairness in user-centric adaptation, advocating for participatory approaches that involve marginalized speech communities in alignment design.

This work is reported in Mire et al. (2025).

4 Task T5.3: Robustness to noisy input (IT*, NAV, UNB)

Proposal

The key highlights from the proposal are listed below.

5.3a hallucinations: for reliable language systems, robustness to noise (e.g., typos, abbreviations and grammatical mistakes in text, background noise in the speech signal, and errors caused by automatic speech recognition in pipeline systems) is required. Current translation systems often hallucinate or produce critical errors in this regime.

5.3b robustness: we will investigate training and adaptation strategies to increase robustness.

Summary of completed work

The original plan was to work on this task from the beginning of the project till month 24. Our contributions are organised under 2 themes, one focused on 5.3a and one on 5.3b (with some natural overlap). In RP1 we focused a bit more on 5.3a. In RP2, we focused on 5.3b.

4.1 Hallucination

Detecting and mitigating hallucinations are key elements in guaranteeing robustness of generation models, and key desiderata for models that may be used for tasks such as real-time or high-risk translation. In RP1, we contributed to theoretical understanding of hallucinations as well as to its effective detection and mitigation. The following three outputs were summarised in D5.1; to date, they have gathered nearly 350 citations (mostly due to the first and third outputs):

- A Comprehensive Study of Hallucinations in Neural Machine Translation (Guerreiro et al., 2023d, see D5.1, Section 4.1.1).
- Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation (Guerreiro et al., 2023b, see D5.1, Section 4.1.2).
- Hallucinations in Large Multilingual Translation Models (Guerreiro et al., 2023a, see D5.1, Section 4.1.3).

4.2 Robustness

ML models can fail in unexpected ways on tested outside their training data distribution. While LLM training has certainly mitigated this, compared to earlier (single task) neural models, the problem remains pervasive across tasks. In RP1, we contributed to improved robustness via variants of system combination in MT decoding and evaluation. The following two outputs were summarised in D5.1; to date, they have gathered nearly 50 citations:

- Translation Hypothesis Ensembling with Large Language Models (Farinhas et al., 2023, see D5.1, Section 4.2.1)

- Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation (Glushkova et al., 2023, see D5.1, Section 4.2.2).

In RP2, we focused on improved robustness via reasoning chains (§4.2.1 and §4.2.2), by data augmentation and alignment using translated data (§4.2.3 and §4.2.4), and we also contribute an analysis of the effects that source-side noise have on the quality of machine translation (§4.2.5).

4.2.1 Improving Multilingual Retrieval-Augmented Language Models through Dialectic Reasoning Argumentations

Retrieval-augmented generation (RAG) grounds large language models (LLMs) in external evidence. Still, it remains brittle when confronted with irrelevant or conflicting sources—problems amplified in multilingual settings where the retrieved knowledge is heterogeneous. The paper proposes Dialectic-RAG (D-RAG), a modular framework that makes RAG-based workflow critical and grounded. D-RAG leads the models to deliver Argumentative Explanations, i.e., a structured reasoning process that systematically evaluates retrieved information by comparing, contrasting, and resolving conflicting perspectives. Given a query and a set of multilingual related documents, D-RAG selects and exemplifies relevant knowledge, filters extraneous content, and clearly provides the final response. Through a series of in-depth experiments, we show the impact of D-RAG both as an in-context learning strategy and for constructing demonstrations to instruct smaller models. The final results demonstrate that D-RAG significantly improves RAG pipelines, requiring low-impact computational effort and providing robustness to knowledge perturbations.

Data and models. We experimented with D-RAG on open-source knowledge-intensive QA benchmarks with multilingual coverage: MLQA, MKQA, XOR-TyDi QA (11 languages across tasks), Natural Questions (English), and BORDERLINES (territorial disputes framed in multiple languages). Wikipedia serves as the knowledge base; documents are retrieved using Cohere's multilingual embeddings (top-5 per query). Models include GPT-4o and the Llama-3 family at 70B, 8B, and 1B scales; decoding is greedy with temperature 0 for determinism. Language-resource disparities in Wikipedia and CommonCrawl are quantified to contextualise retrieval heterogeneity.

Findings. Overall results demonstrate the actual functionality of D-RAG. Indeed, when it is used in an in-context learning (ICL) setting, it delivers clear gains over both non-retrieval baselines and standard RAG on multilingual QA. With GPT-4o, average accuracy increases more than 50% relative to no-RAG and by a further 12.9% over vanilla RAG, while Llama-3-70B exhibits a comparable uplift over RAG (11.9%). By contrast, smaller models struggle to execute the multi-step protocol reliably in ICL settings; however, when fine-tuned on D-RAG demonstrations, they benefit substantially. Specifically, Llama-3-8B surpasses both RAG and supervised fine-tuning on RAG prompts, with reported average gains of roughly 9.6% and 5.5%, respectively. This evidence indicates that dialectic structure yields direct returns for frontier models and, to a considerable extent, functions as an effective supervision signal that provides reasoning benefits to smaller and often more fragile models.

This work is described in (Ranaldi et al., 2025).

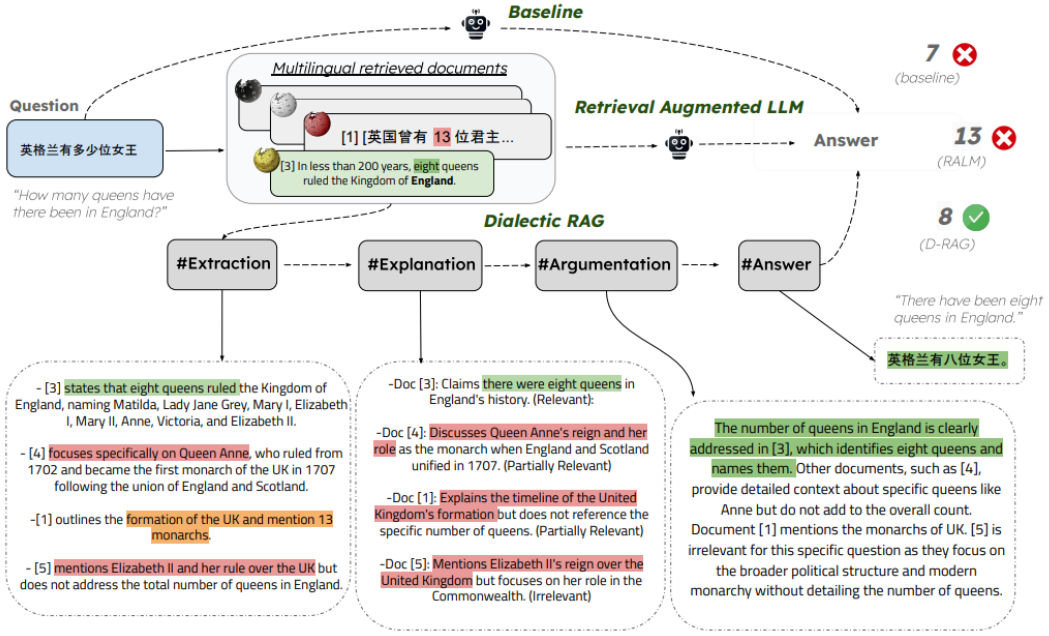


Figure 11: D-RAG allows LLMs to leverage multilingual knowledge-intensive question answering tasks by delivering argumentative explanations that support the final answer.

4.2.2 Empowering Multi-step Reasoning across Languages via Program-Aided Language Models

The paper proposes Cross-PAL, a cross-lingual extension of Program-Aided Language Models and structure reasoning trajectories operating via two modules: an understander that plans a solution in a pivot language (typically English) using program-like steps, and a solver that executes and answers in the original query language. Then, we introduce a self-consistent extension, SCross-PAL, that ensembles multiple cross-lingual reasoning paths to select the most consistent outcome. We show that this method targets an ongoing limitation of multilingual reasoning, such as instability and drop-offs in low-resource languages, by aligning planning in a high-resource language with language-specific execution.

Data and models. The core evaluation centres on multilingual arithmetic reasoning: MGSM (250 GSM8K items translated and double-checked into 10 languages) and MSVAMP (the SVAMP suite in 9 languages). To probe generality beyond arithmetic, we also study results on XCOPA for multilingual causal commonsense reasoning. To deliver broader results, we operate using both close-weight (GPT-based) and open-weight models (Llamas and Phi) of different scales. We compute the final results using the accuracy score computed by exact match against the gold numeral/-text, with answers normalised to the target language format.

Findings. We demonstrate that Cross-PAL improves multilingual reasoning performance across tasks and model sizes. On MGSM and MSVAMP, it achieves higher accuracy than existing prompting methods, with self-consistent Cross-PAL (SCross-PAL) providing further gains. The modular approach is critical, as both single-step prompting and first-step-only variants underperform, highlighting the value of explicit planning followed by execution. Smaller open-weight models

such as Llama-3-8B and Phi-3 benefit substantially, limiting the gap with larger models when led by program-aided demonstrations. We also observe that using English as the pivot language strengthens reasoning, particularly for low-resource languages like Telugu, Swahili, Bangla and Thai, whereas native-language planning is less effective. Self-consistency mode stabilises performance by ensembling across languages, with English playing a pivotal role in low-resource settings. Finally, we prove that the approach generalises beyond arithmetic: on XCOPA, Cross-PAL and its self-consistent variant remain competitive for causal commonsense reasoning, underscoring the applicability of the method.

This work is described in Ranaldi et al. (2024).

4.2.3 Question Translation Training for Better Multilingual Reasoning

Large language models show compelling performance on reasoning tasks but they tend to perform much worse in languages other than English. This is unsurprising given that their training data largely consists of English text and instructions. A typical solution is to translate instruction data into all languages of interest, and then train on the resulting multilingual data, which is called translate-training. This approach not only incurs high cost, but also results in poorly translated data due to the non-standard formatting of mathematical chain-of-thought. In this paper, we explore the benefits of question alignment, where we train the model to translate reasoning questions into English by finetuning on X-English parallel question data. This is shown in Figure 12. In this way we perform targeted, in-domain language alignment which makes best use of English instruction data to unlock the LLMs' multilingual reasoning abilities. Experimental results on LLaMA2-13B show that question alignment leads to consistent improvements over the translate-training approach: an average improvement of 11.3% and 16.1% accuracy across ten languages on the MGSM and MSVAMP multilingual reasoning benchmarks[6].

This work is described in (Zhu et al., 2024b).

4.2.4 The Power of Question Translation Training in Multilingual Reasoning: Broadened Scope and Deepened Insights

This work is an extension of Zhu et al. (2024b), deepening the analysis and insight.

Bridging the significant gap between large language model's English and non-English performance presents a great challenge. While some previous studies attempt to mitigate this gap with translated training data, the recently proposed question alignment framework leverages the model's English expertise to improve multilingual performance with minimum usage of expensive, error-prone translation. In this paper, we explore how broadly this method can be applied by examining its effects in reasoning with and without chain-of-thought, as well as with program-of-thought. We also explore applying this framework to extremely large language models in an efficient manner, such as through proxy-tuning. Experiment results on multilingual reasoning benchmarks MGSM, MSVAMP, XCSQA and XNLI demonstrate that we can extend question alignment framework to boost multilingual performance across diverse reasoning scenarios, model families, and sizes. For instance, when applied to the LLaMA2 models, it brings an average accuracy improvements of 12.2% on MGSM even with the 70B model. To understand the mechanism of its success, we ana-

[6] The project is available at: https://github.com/NJUNLP/QAlign.

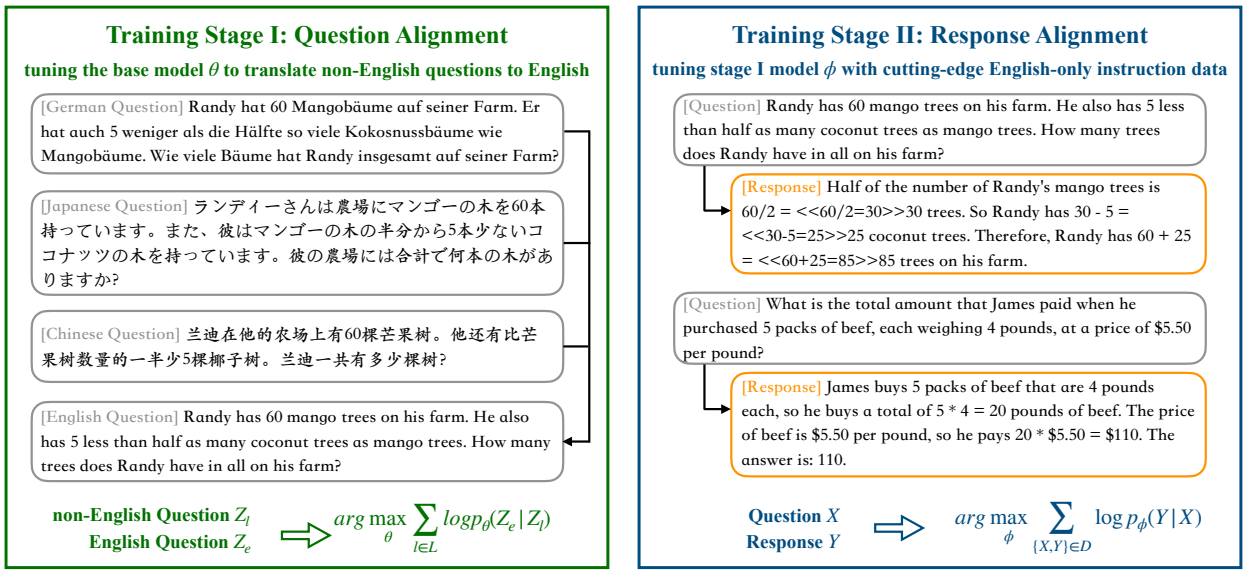


Figure 12: Illustration of our devised two-step training framework. At training stage I (question alignment), we use a set of multilingual questions for translation training. At training stage II (response alignment), we use cutting-edge English-only instruction data for fine-tuning. Due to the established language alignment in stage I, we can utilize LLM's expertise in English to enhance its performance on non-English tasks.

lyze representation space, generated response and data scales, and reveal how question translation training strengthens language alignment within LLMs and shapes their working patterns.

This work is described in (Zhu et al., 2024a).

4.2.5 Did Translation Models Get More Robust Without Anyone Even Noticing?

In this work, we investigate the effect that source-side artificial and natural noise have on the quality of machine translation. Noise has been shown to be a major challenge for conventional NMT models (Belinkov and Bisk, 2018), which are generally trained from scratch on task-specific data, often for a single language pair. We revisit this analysis in light of newer systems based on large multilingual encoder-decoder models and instruction-tuned LLMs, as well as the opaque proprietary system ChatGPT. We explore these questions through experiments on social media text and synthetically noised corpora. These experiments have complementary roles: social media text contains diverse noise phenomena, but isolating their effect is not straightforward because the errors are not labeled. On the other hand, synthetic errors may differ in major ways from "naturally occurring" noise, but they are **interpretable** and **controllable**, offering a way to measure noise *in vitro*. By evaluating on a broad spectrum of error types, we can paint a more vivid picture of what kinds of noise, and at what quantities, cause problems for MT systems.

Our main findings are as follows:

- We show that large pretrained models (specifically, NLLB-200, TowerLLM, and ChatGPT) are much more robust to synthetic source-side errors than conventional single-pair NMT models (see Figure 13), even when their performance is similar on clean data. These results hold across several language pairs and varieties of noise, even though the large models lack architectural features that obviously encourage robustness to character noise.

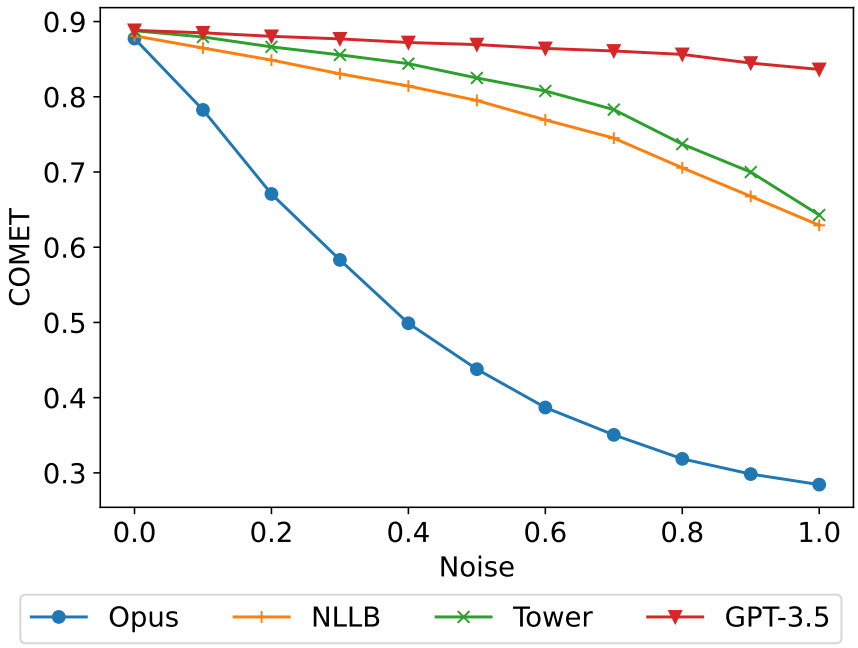


Figure 13: COMET-22 score for English-French on the FLORES-200 devtest set as an increasing proportion of source tokens are noised by randomly swapping a pair of characters.

- We show that models that are robust to synthetic errors also perform better at translating social media text, a variety of "real-world" noise. In addition to conventional reference-based translation experiments, we also include a reference-free setting in which noisy sources can be directly compared to their cleaned counterparts. Both sets of experiments show that LLMs are more robust than conventional models to social media text.
- We show that **source correction pipelines** can be an effective approach to mitigate the impact of synthetic noise without substantially worsening performance on clean data, although they are significantly less effective with stronger models, suggesting that the benefits of source correction and model robustness are not complementary. Source correction is less effective on social media data, likely because there are not enough errors to outweigh the risk of error propagation.

This work is described in (Peters and Martins, 2025).

5 Impact

This work package has produced 41 research papers (17 in T5.1, 14 in T5.2 and 10 in T5.3) across top NLP, CL and ML conferences (*ACL, EMNLP, ICLR, COLM, AISTATS and COLING) and journals (TACL), as well as specialised workshops. Our outputs from RP1 have collectively gathered some 850 citations (220 from T5.1, 230 from T5.2 and 400 from T5.3).[7] We also produced research code bases (approximately one per research paper) serving towards reproducibility of findings but also, in many cases, to support open access to powerful models and tools (e.g., xCOMET, xTower, CREST, MONITOR, ReMaKE, INCLINE, and QAlign, to name a few).

We have co-organised the first two editions of the Uncertainty-Aware NLP workshop https://uncertainlp.github.io, the first (Vázquez et al., 2024) in RP1 (at EACL 2024) and the second in RP2 (at EMNLP 2025). The first edition attracted 30 experts to the programme committee, it received 28 submissions and accepted 12; the second edition attracted 50 experts to the programme committee, it received 46 submissions and accepted 27. This workshop series is a clear extension of our impact and we intended to continue co-organising it.

6 Conclusion

WP5 achieved and arguably surpassed its stated goals on all three tasks. We have contributed datasets, methodology, software and empirical observations to advance various aspects of uncertainty-aware generation, explainability and robustness. Our research output has already been largely cited and is bound to have a lasting impact in the field. We have co-organised two scientific events, helping establish WP5's timely research agenda and also contributing to the project's impact. There were no deviations of action in this package.

[7] We have not collected citations counts for RP2 outputs as those are rather recent.

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.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. Interpreting predictive probabilities: Model confidence or human label variation? In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.24.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Margarida M. Campos, João Cálem, Sophia Sklaviadis, Mário A. T. Figueiredo, and André F. T. Martins. Sparse activations as conformal predictors. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025*, volume 258, pages 2674–2682. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/campos25a.html.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. Generics are puzzling. can language models find the missing piece? 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 6571–6588, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.438/.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. Mgen: Millions of naturally occurring generics in context. In *Society for Computation in Linguistic*, 2025b. doi: 10.7275/scil.3147. URL https://openpublishing.library.umass.edu/scil/article/id/3147/.

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

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. Evaluation of african american language bias in natural language generation. *arXiv preprint*, 2023. URL http://arxiv.org/abs/2305.14291.

Nicholas Deas, Jessica A. Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond U. Patton, and Kathleen McKeown. Phonate: Impact of type-written phonological features of african american language on generative language modeling tasks. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=rXEwxmnGQs.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Bryan Eikema. The effect of generalisation on the inadequacy of the mode. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe, editors, *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 87–92, St Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.uncertainlp-1.9.

Bryan Eikema, Evgenia Ilia, José G. C. de Souza, Chrysoula Zerva, and Wilker Aziz. Teaching language models to faithfully express their uncertainty, 2025a. Preprint, to appear on arXiv.

Bryan Eikema, Anna Rutkiewicz, and Mario Giulianelli. Structure-conditional minimum bayes risk decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, November 2025b. Association for Computational Linguistics. To appear.

António Farinhas, José de Souza, and Andre Martins. An empirical study of translation hypothesis ensembling with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.733. URL https://aclanthology.org/2023.emnlp-main.733.

António Farinhas, Chrysoula Zerva, Dennis Thomas Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=j511LaqEeP.

Dennis Fucci, Marco Gaido, Matteo Negri, Luisa Bentivogli, Andre Martins, and Giuseppe Attanasio. Different speech translation models encode and translate speaker gender differently. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1005–1019, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.78. URL https://aclanthology.org/2025.acl-short.78/.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. What comes next? evaluating uncertainty in neural text generators against human production variability. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.887. URL https://aclanthology.org/2023.emnlp-main.887.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.330. URL https://aclanthology.org/2021.findings-emnlp.330.

Taisiya Glushkova, Chrysoula Zerva, and André F. T. Martins. BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.6.

Gonçalo Emanuel Cavaco Gomes, Bruno Martins, and Chrysoula Zerva. A conformal risk control framework for granular word assessment and uncertainty calibration of clipscore quality estimates. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12348–12365, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.638. URL https://aclanthology.org/2025.findings-acl.638/.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. Investigating african-american vernacular english in transformer-based

text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5877–5883, 2020. URL https://aclanthology.org/2020.emnlp-main.473.

Tobias Groot, Salo Lacunes, and Evgenia Ilia. Learning to vary: Teaching LMs to reproduce human linguistic variability in next-word prediction. In *Second Workshop on Uncertainty-Aware NLP - EMNLP 2025*, 2025. URL https://openreview.net/forum?id=1YGn9OHX7p.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in Large Multilingual Translation Models, March 2023a. URL http://arxiv.org/abs/2303.16104. arXiv:2303.16104 [cs].

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. Optimal transport for unsupervised hallucination detection in neural machine translation. 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 13766–13784, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.770. URL https://aclanthology.org/2023.acl-long.770.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*, 2023c.

Nuno M. Guerreiro, Elena Voita, and André Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia, May 2023d. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.75. URL https://aclanthology.org/2023.eacl-main.75.

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/.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*, 2021.

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. *NeurIPS*, 2021.

Evgenia Ilia and Wilker Aziz. Predict the next word: <humans exhibit uncertainty in this task and language models _____>. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 234–255, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.22.

Evgenia Ilia and Wilker Aziz. Variability need not imply error: The case of adequate but semantically distinct responses, 2024b. URL https://arxiv.org/abs/2412.15683.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1365. URL https://aclanthology.org/P19-1365/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 3, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research (CIFAR), 2009.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *NeurIPS ML Safety Workshop*, 2022.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L. J. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2403.13787.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Youwei Liang, Junfeng He, Gang Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Peizhao Yang, et al. Rich-hf: Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*. IEEE/CVF, 2024. URL https://github.com/google-research-datasets/richhf-18k.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Steven G Luke and Kiel Christianson. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833, 2018.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. 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 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. 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 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. Rejected dialects: Biases against African American language in reward models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7468–7487, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.417. URL https://aclanthology.org/2025.findings-naacl.417/.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *ICLR*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

AI Mistral. Mistral nemo, Jul 2024. URL https://mistral.ai/news/mistral-nemo/.

Tom Mitchell. Twenty newsgroups. uci machine learning repository, 1999.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. Extrinsic evaluation of machine translation metrics. 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 13060–13078, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.730. URL https://aclanthology.org/2023.acl-long.730.

Angelos Nalmpantis, Apostolos Panagiotopoulos, John Gkountouras, Konstantinos Papakostas, and Wilker Aziz. Vision diffmask: Faithful interpretation of vision transformers with differentiable patch masking. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3756–3763, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/XAI4CV/papers/Nalmpantis_Vision_DiffMask_Faithful_Interpretation_of_Vision_Transformers_With_Differentiable_Patch_CVPRW_2023_paper.pdf.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Ben Peters and Andre Martins. Did translation models get more robust without anyone Even noticing? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.122. URL https://aclanthology.org/2025.acl-long.122/.

Suvir Petryk, Diana Chan, Suttida Kachinthaya, Hao Zou, John Canny, Theodora Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. In *North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-main.153/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. Empowering multi-step reasoning across languages via program-aided language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.678. URL https://aclanthology.org/2024.emnlp-main.678/.

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, November 2025. Association for Computational Linguistics. To appear.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. The inside story: Towards better understanding of machine translation neural evaluation metrics. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.94. URL https://aclanthology.org/2023.acl-short.94.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. FOIL it! find one mismatch between image and language caption. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1024. URL https://aclanthology.org/P17-1024/.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. CREST: A joint framework for rationalization and counterfactual text generation. 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 15109–15126, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.842. URL https://aclanthology.org/2023.acl-long.842.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. xTower: A multilingual LLM for explaining and correcting translation errors. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.892. URL https://aclanthology.org/2024.findings-emnlp.892/.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. Control the temperature: Selective sampling for diverse and high-quality LLM outputs. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=lyOC5GCzv4.

Sergey Troshin, Irina Saparina, Antske Fokkens, and Vlad Niculae. Asking a language model for diverse responses. In *Second Workshop on Uncertainty-Aware NLP - EMNLP 2025*, 2025b. URL https://openreview.net/forum?id=Tf73rORM0x.

Dennis Ulmer, Chrysoula Zerva, and Andre Martins. Non-exchangeable conformal language generation with nearest neighbors. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1909–1929, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.129.

Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe, editors. *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, St Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.uncertainlp-1.0.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.155.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Weixuan Wang, Barry Haddow, and Alexandra Birch. Retrieval-augmented multilingual knowledge editing, 2023a. URL https://arxiv.org/abs/2312.13040.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. Assessing the reliability of large language model knowledge, 2023b. URL https://arxiv.org/abs/2310.09820.

Weixuan Wang, Barry Haddow, and Alexandra Birch. Retrieval-augmented multilingual knowledge editing. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.21. URL https://aclanthology.org/2024.acl-long.21/.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. Assessing factual reliability of large language model knowledge. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 805–819, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.46. URL https://aclanthology.org/2024.naacl-long.46/.

Weixuan Wang, Barry Haddow, Wei Peng, and Alexandra Birch. Sharing Matters: Analysing Neurons Across Languages and Tasks in LLMs, 2024c. URL https://arxiv.org/abs/2406.09265.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.270. URL https://aclanthology.org/2025.acl-long.270/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.443. URL https://aclanthology.org/2024.emnlp-main.443/.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. Watching the watchers: Exposing gender disparities in machine translation quality estimation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1228. URL https://aclanthology.org/2025.acl-long.1228/.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. Disentangling uncertainty in machine translation evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.591. URL https://aclanthology.org/2022.emnlp-main.591.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights, 2024a. URL https://arxiv.org/abs/2405.01345.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.498. URL https://aclanthology.org/2024.findings-acl.498/.

Caleb Ziems, William Zhang, Diba Mirza, Maarten Sap, and William Yang Wang. Value: Measuring dialect bias in text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages XXX–XXX, 2022.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D20/D5.2 Final Report on Uncertainty-Aware, Robust and Explainable Models