UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D19 – First Report on Uncertainty-Aware, Robust and Explainable Models

Nature	Report	**Work Package**	WP5
Due Date	28/03/2024	**Submission Date**	dd/mm/2024
Main authors	Wilker Aziz (UVA)		
Co-authors	Barry Haddow (UEDIN), Alexandra Birch (UEDIN), Chrysoula Zerva (IT)		
Reviewers	Laurent Besacier (NAV)		
Keywords	uncertainty quantification, input attribution, hallucination, robustness		
Version Control			
v0.1	**Status**	Draft	22/03/2024
v1.0	**Status**	Final	26/03/2024

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 progress in the first half of the project. WP5 develops techniques to make LLMs confidence-aware, explainable and robust (*e.g.*, to noisy or out-of-domain input), the work is divided in three tasks, each covering one of these themes. The first half of the project saw considerable progress along all three tasks with outputs in the form of peer-reviewed publications, public code and data, as well as co-organisation of events. We did not face nor do we foresee any noteworthy risk, and expect steady progress in the second half of the project.

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, which we cover in detail in sections 2, 3, and 4.

Summary of Output

Manuscripts: 1 journal article (TACL23), 14 conference papers (EMNLP22, ACL23, EACL23, EAMT23, EMNLP23, EACL24, ICLR24), 2 workshop papers (CVPR23, EACL24), and 3 arXiv pre-prints.

Code and data:

- https://github.com/evgeniael/predict_next_word
- https://github.com/dmg-illc/nlg-uncertainty-probes
- https://github.com/Vicky-Wil/MONITOR
- https://github.com/Vicky-Wil/ReMaKE
- https://github.com/deep-spin/uncertainties_MT_eval
- https://github.com/Kaleidophon/non-exchangeable-conformal-language-generation
- https://github.com/deep-spin/non-exchangeable-crc
- https://github.com/AngelosNal/Vision-DiffMask
- https://github.com/deep-spin/crest/
- https://github.com/Unbabel/COMET/tree/explainable-metrics
- https://huggingface.co/collections/Unbabel/xcomet-659eca973b3be2ae4ac023bb
- https://github.com/deep-spin/hallucinations-in-nmt
- https://github.com/deep-spin/ot-hallucination-detection
- https://github.com/deep-spin/lmt_hallucinations
- https://github.com/deep-spin/translation-hypothesis-ensembling
- https://github.com/deep-spin/robust_MT_evaluation

Events: 1 workshop (UncertaiNLP) at EACL24.

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. Indeed, the first half of the project saw considerable progress along the two dimensions above (*i.e.*, uncertainty-awareness and interpretable uncertainty).

Since the initial proposal, this task saw no decrease in relevance, on the contrary, uncertainty-aware methods are growing in importance and popularity (see Figure 1).[1] Next, we report on our progress

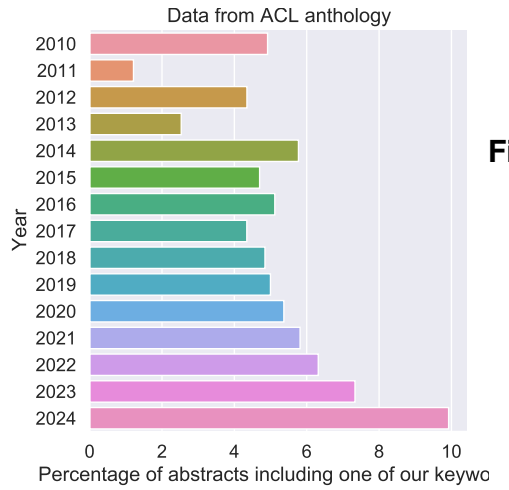

Figure 1: The plot shows the increase in number of abstracts (from the ACL anthology https://aclanthology.org) including an uncertainty-related keyword. Image from the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP; Vázquez et al., 2024), which UTTER co-organised and supported.

in the first 18 months of the project. Our contributions to this task are organised under 4 subthemes: 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.

[1] Uncertainty-related keywords: uncertainty, variability, variation, aleatoric, epistemic, evidential, confidence, disagreement, multiple perspectives, multiple judgements, multiple annotation, multiple annotators, multiple references, active learning, calibration, conformal prediction, probabilistic inference, approximate inference, Bayesian inference, BSL, Bayesian DL, Bayesian deep learning, BNN, Bayesian neural net, sampling, decision-making, utility-aware, controllable generation, selective prediction, selective generation, statistical evaluation.

2.1 Statistical Evaluation of Natural Language Generators

Natural language generators are built upon probabilistic models which inherently pack a highly structured representation of uncertainty about responses given a prompt. The next three pieces of work advance statistical evaluation of these objects, which is a crucial ingredient in diagnosing and driving progress. These contributions advance sub-goal 5.1a (uncertainty-awareness).

2.1.1 Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?

In common language, uncertainty refers to "a state of not being definitely known or perfectly clear; a state of doubt".[2] In statistics and machine learning, uncertainty is taken as a state to be represented (Lindley, 2013; Halpern, 2017)—the state of the world as a function of inherently stochastic experiments or the state of knowledge of an agent observing or interacting with the world—and its mathematical representation requires prescribing a probability measure (Kolmogorov, 1960). In modern NLP, neural networks are the de-facto standard to predict complex probability measures from available context (Goldberg and Hirst, 2017): given an input (or prompt), a neural network prescribes a representation of uncertainty over the space of responses (*e.g.*, strings or classes), typically, by mapping the input to the parameter of a probability mass function (*e.g.*, in text classification, inputs are mapped to the probability masses of each outcome in the label space).

Recently, transformer-based large language models (LLMs) are becoming increasingly powerful and display remarkable abilities on complex classification tasks, leading to an increased deployment in user-facing applications. This motivates the need for models that can signal when they are likely to be wrong (an aspect of trustworthiness), and models that can capture different linguistic and human interpretations (an aspect of language including fairness). We identify two perspectives in the literature whereby the exact same representation of uncertainty—the predictive distribution over outcomes—is sometimes interpreted as an indication of confidence in model predictions (**P1;** Desai and Durrett, 2020; Dan and Roth, 2021; Jiang et al., 2021) and other times as an indication of variation in human perspectives (**P2;** Plank, 2022).

Our position paper (Baan et al., 2024) provides clarity and accelerates progress by:

- Identifying these two perspectives on the predictive distribution and examining how each evaluates the quality of predictive distribution.
- Documenting their merits and limitations, and relating them to popular notions of *aleatoric* and *epistemic* uncertainty.
- Taking the position that both perspectives contribute to trustworthy and fair NLP systems, but that exploiting a single predictive distribution is limiting, and highlighting exciting directions towards models that can predict distributions over human or linguistic interpretations, and simultaneously abstain from answering when lacking such knowledge or skills.

This work is reported in (Baan et al., 2024).

[2] Oxford English Dictionary, accessed October 13th 2023.

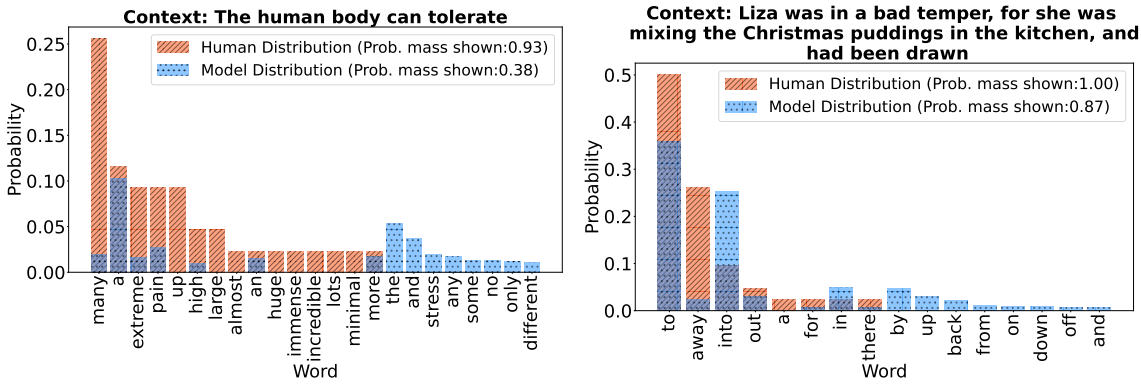

Figure 2: Estimated human and model distributions for two example contexts (15 most probable words of each distribution).

2.1.2 Predict the Next Word: ⟨*Humans exhibit uncertainty in this task and language models* _____⟩

Autoregressive LMs are trained to assign probability to the next token conditional on a context. Their conditional predictive distributions can be viewed as a representation of uncertainty of human production variability – specifically, one that reflects the production variability of the population that generated their training data. Despite how natural plausible variability is, LMs are not consistently subjected to such variability during their training. This led us to question and hence investigate their ability to predict it well.

In order to assess whether a language model can serve as a good proxy to the production variability humans exhibit, we exploit the following desideratum, previously termed as calibration to human uncertainty (Baan et al., 2022) in the context of text classification: a language model should assign probability to any next word candidate similar to the proportion of humans assigning that word as the next one. We are able to appreciate production variability in humans, by providing a group of humans with a context and asking each of them to provide which word they think follows. We can prompt the model we'd like to assess using the same context and sample word continuations. Using the human and model samples, we obtain Monte Carlo estimates of the conditional probability distributions over word continuations given a prefix (see Figure 2 for some examples); which we can compare using Total Variation Distance (TVD). For the purposes of our analysis, we use Provo Corpus (Luke and Christianson, 2018), a dataset of around 2.6k contexts with, on average, 40 human word continuations per context. We assess three pre-trained LMs (GPT2 (Radford et al., 2019), Bloom (Scao et al., 2022), ChatGPT (OpenAI, 2022)) representative of different model sizes and training objectives. We assess them by plotting distribution of TVD values across all contexts and computing the average TVD. Our contributions can be summarised as follows:

- We introduce a technique to assess model's calibration against human uncertainty, our technique does not require open-access to the model (that is, it is sufficient to have access to an inference-only API) and is applicable regardless of tokenisation strategy.
- The models we analysed do not reliably reproduce the range of variability recorded in Provos.
- We analyse potential predictors of this inability and find that models particularly struggle when predicting words following contexts that allow for higher plausible variability.

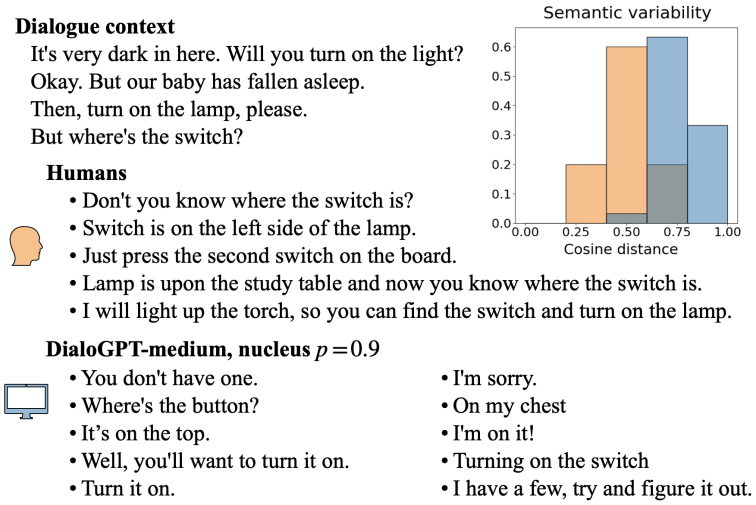

Figure 3: Production variability observed in 5 human responses vs 10 responses generated by DialoGPT. The graph presents the distribution of pairwise cosine distances: generated responses exhibit higher semantic variability than human responses. The generator's semantic uncertainty is too high in this dialogue context.

We believe that models explicitly supervised with variability, might be able to better capture human production variability, which we leave as future work.

This work is reported in (Ilia and Aziz, 2024).

2.1.3 What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability

Humans display great variability in language production, in particular when the context or the task are open-ended, such as in storytelling or in dialogue. Given a story prompt, for example, there are many plausible ways in which different humans (or a single writer, if asked multiple times) may tell the story (Fan et al., 2018). We refer to this phenomenon as *production variability*. Production variability in humans has two main sources. First, when situated in a context, speakers may entertain variable communicative goals (Searle, 1969; Sacks et al., 1974; Austin, 1975), and the number and variety of plausible communicative goals depends on the production task (Jokinen, 1996). Translation, for instance, defines the communicative goal almost unequivocally while a dialogue context might allow for a wide variety of communicative goals (expressed, *e.g.*, as a request, an assertion, or a yes-no question). The second source of variability is the fact that even when context and communicative goal are fixed, speakers' linguistic realisations of the communicative goal may vary (Levelt, 1993). Both sources of variability apply to individuals as well as to populations: if an expert is asked to simplify a complicated sentence multiple times, they may perform different rewriting transformations (*e.g.*, paraphrasing, reordering, or sentence splitting) and produce different texts (Alva-Manchego et al., 2021); the same is true if multiple experts are asked to perform a task (Xu et al., 2015). If we are to regard a Natural Language Generation (NLG) system (or *text generator*) as a good model of human production, it should capture the variability observed in humans.

Text generators combine two mechanisms: (i) an underlying statistical model—typically, an autoregressive factorisation of the probability of sequences, with conditional token probabilities predicted

by a neural network; and (ii) an iterative decoding algorithm that chains samples from next token distributions into a complete production. Together these two mechanisms specify a probability distribution over sequences of tokens, which can be regarded as a representation of the model's uncertainty about productions for a given generation context. In this work, we assess whether this representation of uncertainty is in compliance with production variability exhibited by a population of humans—which in turn, we argue, can be regarded as an expression of *aleatoric* uncertainty, *i.e.*, irreducible uncertainty due to the stochastic nature of the data generating process (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021). In other words, we compare the distribution over productions of a text generator against the distribution over the productions of a population of human speakers, given the same context (see Figure 3 for an overview).

Quantifying the closeness in distribution between a text generator and a human population is difficult: we only have an iterative view into the generator's distribution; the 'human distribution' is an implicit or even hypothetical object; and in both cases, the sample space is large or even unbounded. We can, however, compare these two objects via the samples they produce and assess their statistical distance—which is what we propose here. For each individual generation context, we compare scalar properties of generations (through repeated model sampling) and human productions (using multi-reference NLG datasets). In particular, we *probe* for lexical, syntactic, and semantic distance between productions, thus allowing for a quantitative and interpretable assessment of uncertainty.

Our contributions are summarised as follows:

- We find that the uncertainty of neural text generators is higher than justified by human production variability in open-ended tasks, like story generation and open-domain dialogue; and that it is lower on more constrained tasks, like machine translation and text simplification (see Figure 4).
- Popular decoding algorithms, which bias away from the distribution of the generator's underlying statistical model (*e.g.*, top-k, top-p, or locally typical, rather than ancestral sampling), have a limited impact on the generator's ability to faithfully represent human variability.
- We complement our quantitative assessments with a detailed analysis of individual generation contexts, which sheds light on whether a generator has robustly learned to reproduce degrees and aspects of human variability plausible for the communicative task.

Beyond the experimental results obtained on our selection of models and tasks, our work has important implications for NLG evaluation and data collection. Multiple samples and, when possible, multiple references, should be used to assess the statistical fit of text generators. Our approach, complementary to other types of automatic evaluation, makes model assessments particularly insightful and trustworthy because it does not judge a model only by a single output but also, intuitively, by *what it could have generated*—and it does so for each individual input in the test set. We therefore hope our framework will be used by the community as an evaluation criterion for NLG systems, especially to assess them in more open-ended tasks.

This work is reported in (Giulianelli et al., 2023).

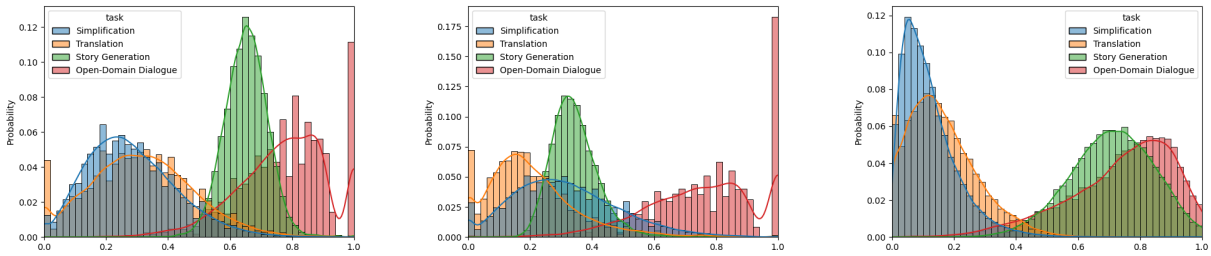

Figure 4: Human production variability across four NLG tasks. The values on the horizontal axis are single samples of lexical (unigram), syntactic (POS bigram), or semantic (cosine) distance between two randomly sampled productions for each input. Re-sampling productions results in nearly identical marginal distributions. Probability mass on the right side signals high distance and thus high variability, and vice versa.

2.2 Improved Text Generation (Decoding)

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 the next two pieces of work, we advance such decoders: we discuss the limits of existing explanations for certain degeneracies common in text generation and contribute a new hypothesis, and combine the uncertainty representation by different models into an improved contrastive decoder. These contributions advance sub-goal 5.1a (uncertainty-awareness).

2.2.1 The Effect of Generalisation on the Inadequacy of the Mode

The highest probability sequences of most natural language generation models tend to be degenerate in some way, a problem known as the inadequacy of the mode (Eikema and Aziz, 2020). While many approaches to tackling particular aspects of the problem exist, such as dealing with too short sequences (Wu et al., 2016) or excessive repetitions (Xu et al., 2022), explanations of why it occurs in the first place are rarer and do not agree with each other (Meister et al., 2022; Yoshida et al., 2023). We challenge the existing hypotheses, arguing they paint an incomplete picture and bring light to the role that generalisation may play in causing the inadequacy of the mode. We argue that in order to generalise well, a neural network needs to map different local contexts to similar continuous representations. As a result, their observations jointly contribute to the same next-word distributions. We hypothesise that this clustering is not always due to actual linguistic equivalence, i.e. human continuations for those contexts may distribute differently. As a result, the introduced spread may lead to inadequate modes for some contexts, which may also result in inadequate modes when compounded to the sequence-level distributions. Therefore, we argue that inadequate modes in natural language generation models may be an inevitable consequence of our desire to generalise to unseen contexts under the current modelling and data constraints.

This work is reported in (Eikema, 2024).

2.2.2 Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models

Hallucinations are a rare but problematic phenomenon in NMT (Neural Machine Translation) whereby the target side output is repetitive or fluent but not grounded in the source sentence (Ji et al., 2023). Even though hallucinations are rare in NMT, they are a significant problem as they undermine trust in deployed NMT systems. Hallucinations occur when the target side sentence is detached from the source side sentence (Wang and Sennrich, 2020; Dale et al., 2023), or in other words, when there is a low contribution of the source sentence to the generation of the target sentence.

Previous work on mitigating hallucinations has focused on sampling translations and reranking them according to quality metrics (Dale et al., 2023; Guerreiro et al., 2023e). Separate to this, Li et al. (2022) proposed Contrastive Decoding (CD) as a way of mitigating bad behaviour (such as excessive repetition and low diversity) when generating from unconditional language models. CD is a decoding algorithm that maximises the difference between the log probabilities of a strong expert and a weak amateur model (equivalent to maximising the ratio of probabilities). A threshold is applied so that decoding follows the expert when it is more confident. The intuition behind CD is that the amateur model is more prone to certain types of low-quality generation, so by subtracting the log probabilities, these are removed. We hypothesise that by using CD with an amateur, which is prone to source detachment, we can mitigate hallucinations in NMT.

In order to create an amateur with low source attachment, we experiment with different strategies for reducing the role of cross-attention. The simplest is the no-encoder strategy, where the amateur is a decoder-only version of the expert. In our other strategies, we retain the encoder and cross-attention but impose uniform attention, remove attention from the most highly attended source position, or scale down all cross-attention values.

In contrast to unconditional generation, NMT should be more strictly grounded in the source sentence. Additionally, hallucinations only account for a small proportion of translations, and hence, mitigation of hallucinations must not come at the cost of reduced performance on other sentences. As such, increasing the diversity of the translations is less desirable than it is in unconditional generation. Ideally, CD would only take effect when a model is hallucinating. To address this issue, we experiment with a novel variant of CD that dynamically adjusts the subtraction's magnitude based on a distribution's maximum value.

We evaluate our approach on large multilingual models, which have recently been shown to be prone to hallucinations (Guerreiro et al., 2023a). Specifically, we use the M2M family of models (Fan et al., 2020) and consider the 418M (Small) and 1.2B (Medium) versions.

We summarise our contributions as follows:

- We show that using CD in conjunction with amateur models that have reduced source contributions mitigates hallucinations.
- We extend the CD algorithm, dynamically setting the weight given to the amateur to limit the effect of CD when the expert is confident.
- We evaluate across 21 language pairs using the M2M family of models on the FLORES-101 dataset, reporting a mean increase of 14.6 ± 0.5 and 11.0 ± 0.6 COMET on sentences causing hallucinations for the Small model and Medium models respectively.

This work is reported in (Waldendorf et al., 2024).

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. In the next two pieces of work, we develop data and methodology to assess the parametric memory of LLMs as well as to edit this memory in order to fix incorrectly stored facts. These contributions advance sub-goal 5.1a (uncertainty-awareness).

2.3.1 Assessing the Reliability of Large Language Model Knowledge

Recently, large pre-trained language models (LLMs) have been used as de facto storage for factual knowledge (Petroni et al., 2019). However, applying LLMs to real-world scenarios inevitably leads to language generation deviating from known facts (aka "factual hallucination" (Chang et al., 2023)) due to multiple causes. For example, Cao et al. (2021a) argued that the performance of an LLM is over-estimated due to biased prompts over-fitting datasets (also referred to as the framing effect in Jones and Steinhardt (2022)) and in-context information leakage. Given the variability of LLMs' performance under different prompts and contexts, it becomes evident that relying solely on accuracy as an evaluation metric is insufficient. We also need to gauge how robust LLMs are to variations in prompting.

In short, we identify two issues with the evaluation of factual knowledge in LLMs using accuracy:

Prompt framing effect: An LLM generates different predictions depending on how prompts are framed. Predictions are associated with prompts instead of factual knowledge learned in LLMs. **Effect of in-context interference:** An LLM leverages in-context information during its decoding stage, but this information may negatively affect an LLM's prediction during knowledge probing. In other words, a model may be influenced to give an incorrect response by information in its context.

We argue that a metric of factual accuracy for LLMs should take into account not just top-1 performance, but should consider the probability that the model assigns to the correct answer. The metric should also consider how the model responds under a variety of prompts. To address these we propose a novel distance-based approach **MOdel kNowledge relIabiliTy scORe (MONITOR)** which captures the deviation of output probability distributions under contexts of prompting variance, interference from mispriming (Kassner and Schütze, 2020) and positively-primed prompts. MONITOR is designed to compute the distance between the probability distributions of a valid output and its counterparts produced by the same LLM probing the same fact using different styles of prompts and contexts. Experiments on a comprehensive range of 12 LLMs demonstrate the effectiveness of MONITOR in evaluating the factual reliability of LLMs while maintaining a low computational overhead.

The contributions of this work are:

1. We propose a novel method to assess the factual reliability of LLMs in the presence of the prompt framing effect and in-context interference. The proposed metric, MONITOR, can be used in conjunction with an end-to-end metric (i.e., accuracy) as part of a multi-dimensional approach to LLM knowledge evaluation.

2. We construct the **FKTC** (Factual Knowledge Test Corpus) by developing question answering probing prompts (210,171 prompts in total) based on 16,167 triplets of 20 fact datasets from T-REx corpus (Elsahar et al., 2018). We will release **FKTC** to the public to foster research works along this line.

This work is reported in (Wang et al., 2023c).

2.3.2 Retrieval-augmented Multilingual Knowledge Editing

Large Language Models (LLMs) are being used as sources of factual knowledge for search engines and other downstream tasks. Despite their considerable progress, knowledge generated by LLMs can be incorrect or become obsolete in a changing world. Pre-training from scratch or fine-tuning LLMs to adapt them to new knowledge is computationally expensive and not guaranteed to work. Knowledge editing (KE) methods (Zhu et al., 2020; Cao et al., 2021b) have been proposed as effective and economic alternatives to fine-tuning when specific factual knowledge needs to be added or updated. KE involves either updating the parameters of a model (Dai et al., 2022a; Mitchell et al., 2022a; Meng et al., 2022, 2023; Dai et al., 2022b) or adding extra components to an LLM (Mitchell et al., 2022b; Zheng et al., 2023). For example, KE can be used to correct the answer to this question *"Who is the foreign secretary of the UK?"* from *"James Cleverly"* (true until mid November 2023) to *"David Cameron"*, who has recently been appointed to the post.

Despite significant interest in this problem, current research on KE predominantly concentrates on a monolingual setting, where both the injected knowledge and the subsequent queries to the LLM are in English (Mitchell et al., 2022a; Meng et al., 2022, 2023; Mitchell et al., 2022b; Zheng et al., 2023). Companies serving a multilingual customer base need to consider the multilingual KE case, where KE is done in one language and this propagates to queries and answers in all other languages. While Wang et al. (2023a) explored the cross-lingual applicability of knowledge editing to the English-Chinese cross-lingual scenario, their primary focus was to highlight the challenges rather than develop a functional KE approach in a multilingual setting.

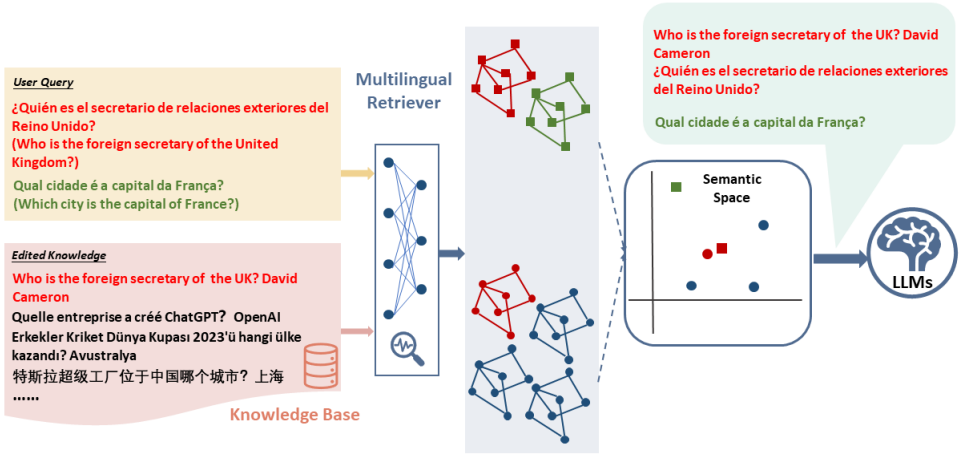

Figure 5: ReMaKE attaches in-context knowledge to an LLM prompt when it is retrieved (red example where the edited knowledge is in English and a user query is in Spanish) from a customer-defined multilingual knowledge base. When no edited knowledge is retrieved (green example) the prompt is passed to the LLM unchanged.

Drawing inspiration from in-context learning (ICL), in-context knowledge editing (IKE) uses prompts to edit factual knowledge. It is noted that IKE is so far the only method demonstrating positive results in the cross-lingual KE task setting (Wang et al., 2023a). However, IKE requires explicit provision of new knowledge every time an LLM is used, confining its practicality and scalability in real-world applications. In addition, IKE suffers when irrelevant facts are included in the prompt (Wang et al., 2023c) especially in scenarios where a substantial number of facts are being edited.

In this paper, we propose **Retrieval-Augmented Multilingual Knowledge Editor (ReMaKE)** that integrates multilingual retrieval from a knowledge base with in-context learning. ReMaKE concatenates the retrieved knowledge from an external database with a user query to create the prompt. The proposed multilingual retriever grounds the ReMaKE to the retrieved accurate and up-to-date information highly relevant to user queries, therefore effectively mitigating the contextual interference due to irrelevant context. In this way, the generated prompts are able to guide the LLMs in producing accurate responses associated with the injected knowledge. ReMaKE leverages a knowledge base's ability to scale to further enhance IKE's knowledge editing performance in real-world application scenarios where large volumes of edits are in scope. Figure 5 shows the architecture of the proposed retrieval-augmented multilingual knowledge editor. Our main contributions are listed below:

- **Multilingual knowledge editing**: ReMaKE extends the scope of knowledge editing practices across language boundaries. Given that the multilingual knowledge base and multilingual retriever operate independently to a specific LLM, ReMaKE is a **plug-and-play** tool applicable to any LLM. It is **scalable**, capable of editing a large number of knowledge. Experiments show ReMaKE outperforms baseline methods by a significant margin in the average accuracy score (up to +40.53%).
- **Multilingual editing dataset**: We build a machine-translated multilingual knowledge editing dataset (**MzsRE**) in 12 languages: English, Czech, German, Dutch, Spanish, French, Portuguese, Russian, Thai, Turkish, Vietnamese, and Chinese using the zsRE testset (Levy et al., 2017). The dataset will be made available to the community.

This work is reported in (Wang et al., 2023b).

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 the next three pieces, we contribute towards more human interpretable forms of uncertainty quantification, for example by disentangling uncertainty representations along aleatoric and epistemic dimensions, and by creating so-called (conformal) prediction sets (roughly speaking, a generalisation of confidence intervals that works in continuous and discrete spaces alike). These contributions advance sub-goal 5.1b (interpretable uncertainty).

2.4.1 Disentangling Uncertainty in Machine Translation Evaluation

In this work, we focus on uncertainty estimation over the predictions of machine translation (MT) evaluation metrics. We were specifically interested in departing from previous work (Glushkova et al., 2021), that explored sampling-based methods for uncertainty quantification such as Monte Carlo dropout and deep ensembles and exploring powerful uncertainty quantification methods that can differentiate between aleatoric (data) and epistemic (model) uncertainties. We compare the following methods:

1. **Heteroscedastic Regression**: This approach models aleatoric (data) uncertainty as observation noise, under the assumption that observed aleatoric uncertainty varies across instances, influenced by factors such as noisy references or inconsistent human annotations. We thus train a COMET Rei et al. (2020) model using heteroscedastic regression, to predict not only a quality score for each instance but also a variance estimate σ^2 for this score, serving as a measure of aleatoric uncertainty.

2. **Divergence Minimization**: In this case, we leverage multiple annotations for each training instance, using annotator disagreement as a proxy for data uncertainty. We then train a model using a divergence minimization objective, specifically the Kullback-Leibler (KL) divergence between the distribution of annotator scores and the predicted distribution of quality scores.

3. **Direct Uncertainty Prediction (DUP)**: Contrary to the methods focusing on aleatoric uncertainty, DUP targets epistemic (model) uncertainty (e.g. uncertainty due to out-of-domain data or unseen linguistic constructions). DUP, inspired by Jain et al. (2021) treats the total uncertainty as an approximation of the model's generalization error, directly learning to predict uncertainty based on observed prediction errors. Thus, we can train training a model for DUP involves a two-step process: first, generating quality score predictions for a given dataset using an initial model; and second, training a separate DUP model to predict the uncertainty of these predictions by estimating the error between the predicted scores and the true human judgments.

Our experiments demonstrate that the proposed methods significantly improve uncertainty prediction in MT evaluation, obtaining better correlations with error on the WMT metrics task datasets, as well as outperforming previous methods on a range of other performance indicators. More importantly, we show that methods targeting different uncertainties can better address specific tasks. As shown in Figure 6 heteroscedastic regression and divergence minimization approaches showed a marked ability to capture aleatoric uncertainty, effectively identifying instances with low-quality references. Instead, DUP proved more adept at reflecting increased uncertainty for out-of-domain data, highlighting its potential for enhancing model reliability in diverse evaluation scenarios (see Figure 7).

This work is reported in (Zerva et al., 2022).

2.4.2 Non-Exchangeable Conformal Language Generation with Nearest Neighbours

Extending our work on uncertainty we explored conformal prediction (Angelopoulos and Bates, 2021; Papadopoulos, 2008; Vovk et al., 2005), a framework that allows us to quantify the uncertainty of predictions, under established statistical guarantees. It provides a way to generate

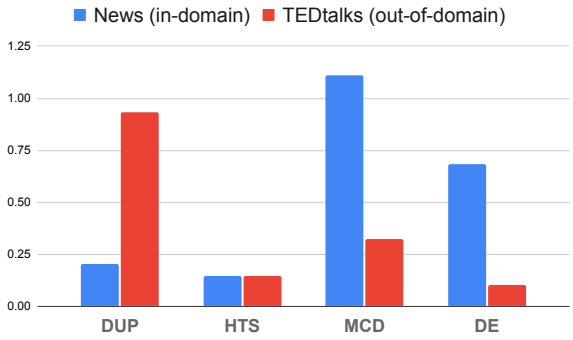

Figure 6: Average uncertainty on two En-Ru test sets: in-domain (News) and out-of-domain (TED talks). Our proposed method that handles epistemic uncertainty (DUP) exhibits higher uncertainty on the out-of-domain dataset

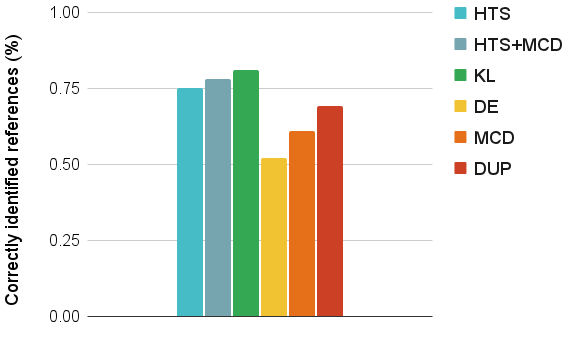

Figure 7: Percentage of correctly recognized references with higher quality by different uncertainty predictors. HTS and KL methods that target aleatoric uncertainty exhibit significantly better performance.

prediction intervals (or sets) that come with a guarantee on coverage, i.e., the expectation that they contain the true value with some probability, under the assumption that the data are exchangeable. While it is an appealing framework, without the need for underlying assumptions on the data or output distributions, the exchangeability assumption poses a challenge for sequence generation tasks such as MT, since there is a dependency on previously generated tokens. As such, we turn to extensions for non-exchangeable data, proposing novel frameworks for non-exchangeable conformal prediction and non-exchangeable conformal risk control that can be used for NLP tasks that violate the exchangeability constraint.

We introduce a novel method based on *non-exchangeable conformal prediction* (Barber et al., 2023) and to apply it to language generation to produce calibrated prediction sets. Our approach leverages advancements in non-exchangeable conformal prediction to generate calibration sets dynamically during inference. We specifically propose to use a *k nearest neighbours* (kNN) method to estimate the most relevant data points from the calibration set at each generation step. This allows for token-level calibrated prediction sets that adapt to the current generation context without necessitating additional model training.

Our method's efficacy is validated across language modelling and machine translation tasks, demonstrating its capability to produce tighter prediction sets with improved coverage compared to previous methods (Ravfogel et al., 2023). We show that using our estimated prediction sets for sampling during generation (non-X conformal sampling), our approach maintains or even enhances generation quality, compared to commonly used sampling methods such as nucleus sampling or top-k sampling. Furthermore, Figure 8 showcases our approach's ability to maintain desired coverage levels even under conditions of distributional shift, highlighting its robustness and adaptability.

Our findings underscore the potential of non-exchangeable conformal sampling to offer a theoretically principled way to sample from language models with conformal guarantees. This provides a more reliable and interpretable framework for evaluating and generating text across various applications.

This work is reported in (Ulmer et al., 2024).

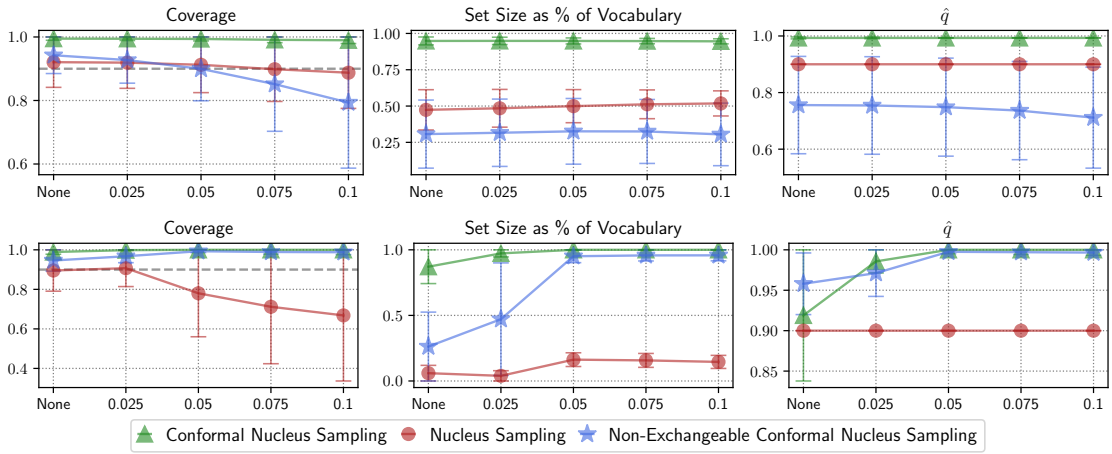

Figure 8: Coverage, average set size and $\hat{q}$ based on the noise level on the de → en MT task (top) and open text generation task (bottom). Error bars show one standard deviation.

2.4.3 Non-Exchangeable Conformal Risk Control

The previously presented work accounts for setups where we are interested in calibrating with respect to coverage but does not account for scenarios where we want to control for a different function. This work extends lines of research in non-exchangeable conformal prediction and focuses on non-exchangeable conformal risk control (Non-X CRC). Our framework is designed to control the expected value of *any monotone loss function* in scenarios where data exchangeability is violated. This extension is crucial for real-world applications where data often exhibits sequential dependencies (e.g. time series), change points, or other forms of distribution drift.

Non-X CRC is characterized by its flexibility and minimal assumptions. It allows for data weighting based on relevance to a given test example, facilitating tighter bounds on expected loss. Our empirical studies underscore the utility of non-X CRC across various tasks: multilabel classification with synthetic data, monitoring electricity usage, and open-domain question answering. These experiments highlight the method's capability to minimize specific loss functions such as the false negative rate and λ-insensitive absolute loss, and to bound the best F1-score, showcasing its versatility and effectiveness. A key finding is the adaptability of non-X CRC in maintaining calibrated prediction sets even under significant data distribution changes. For instance, our framework demonstrated superior performance in adjusting to distribution drift and change points in synthetic time series data, outperforming standard conformal risk control methods.

Specifically for the case of open-domain question answering, which is more relevant to the UTTER project, the task involves generating accurate answers to factoid questions based on a large collection of documents. We controlled for the best token-based F1-score of the prediction set over all pairs of predictions and answers. We employed dynamically computed weights determined by the semantic similarity between the calibration questions and the test question.

The main findings from our experiments on the Natural Questions dataset highlight the efficacy of non-X CRC compared to a CRC baseline as shown in Figure 9. Specifically, we observed that while the overall test risk was comparable between our method and the standard CRC approach, the prediction sets generated by non-X CRC were consistently smaller. This indicates that by incorporating a nuanced weighting scheme based on semantic similarity, non-X CRC is able to achieve the desired risk level with more concise and potentially more accurate prediction sets.

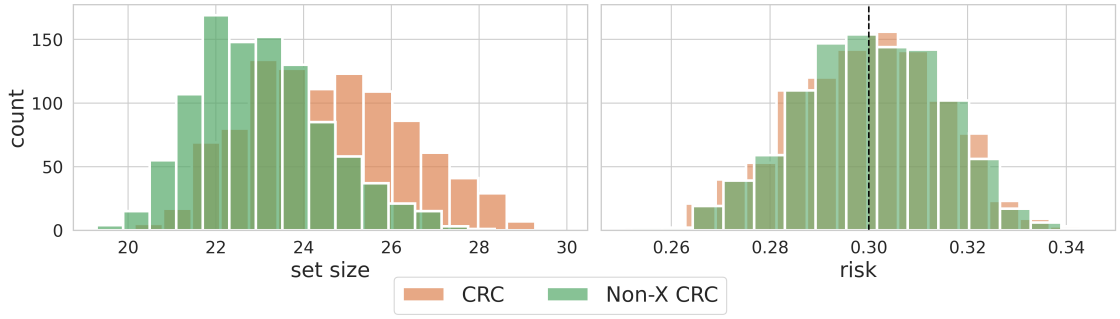

Figure 9: F_1-score control on the Natural Questions dataset. Average set size (left) and risk (right) over 1000 independent random data splits.

These results underscore the potential of non-X CRC to enhance the precision and reliability of open-domain QA systems, that can be extended to other NLP tasks and loss functions.

This work is reported in (Farinhas et al., 2024).

Plans for future work

This task remains very relevant for UTTER's goals. In the second half of the project, we intend to look into more ideas for tuning (*e.g.*, fine tuning or instruction tuning) LLMs to better represent uncertainty, to represent it and/or communicate it in human-readable ways, and to make better use of this representation (*e.g.*, in decoding, to detect or mitigate hallucinations, to deliver robustness to noise and ambiguity, etc.). To reflect the growing impact of uncertainty in LLMs (from design to application), unlike what we stated in the original plan, this task may extend until the end of the project.

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. Indeed,

the first half of the project has already seen considerable progress along the two dimensions above (*i.e.*, explaining predictions and transparent evaluation).

Next, we report on our progress in the first 18 months of the project. 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 in terms of input attribution methods, in the next two pieces we develop faithful attributions for vision Transformers and for explaining and improving text generation. These contributions advance sub-goal 5.2a (explaining predictions).

3.1.1 Vision DiffMask: Interpretability of Computer Vision models with Differentiable Patch Masking

Transformers (Vaswani et al., 2017) have led to various breakthroughs in NLP, but also, more recently, in computer vision (CV) with the Vision Transformer (ViT; Dosovitskiy et al., 2021). These models are able to extract complex features, which have pushed significant improvements in performance on classic tasks (*i.e.*, image classification) and opened the path to more complex ones (*e.g.*, image captioning and other multimodal applications). Despite their success, ViT, as any other Transformer, is an opaque model component that lacks interpretability.

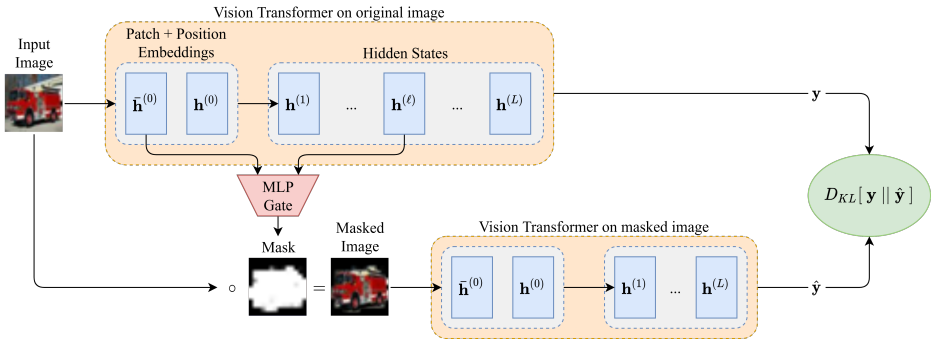

Figure 10: Overview of Vision DiffMask's architecture. Adapted from De Cao et al. (2020).

Explanations for Transformers typically come in the form of saliency maps over the input image, and are often computed with *gradient-based* methods (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2017) or *attribution propagation* methods Binder et al. (2016); Shrikumar et al. (2017); Chefer et al. (2021). However, these approaches can not guarantee that (**a**) the model is fully ignoring low-scored features, or (**b**) the model's output distribution is preserved in their absence. In our earlier work (De Cao et al., 2020) in NLP, we presented DiffMask, a novel method to predict attribution masks over the input text when conditioned on the hidden representations of a language model. DiffMaskis a trainable probe, optimised to keep the minimal subset of the input that produces a similar output distribution over the target labels.

We extend DiffMask to the vision domain by introducing Vision DiffMask, an interpretation network that predicts saliency maps for models following the ViT's architecture (see Figure 10). Our method uses a gating mechanism on each of ViT's layers. During training, all gates cast a binary vote on whether to keep each of the input patches. The votes are aggregated across layers, through

element-wise multiplication, to obtain a single mask to apply at the input. During inference, each gate predicts a probability instead of a binary vote, leading to an attribution map over the input.

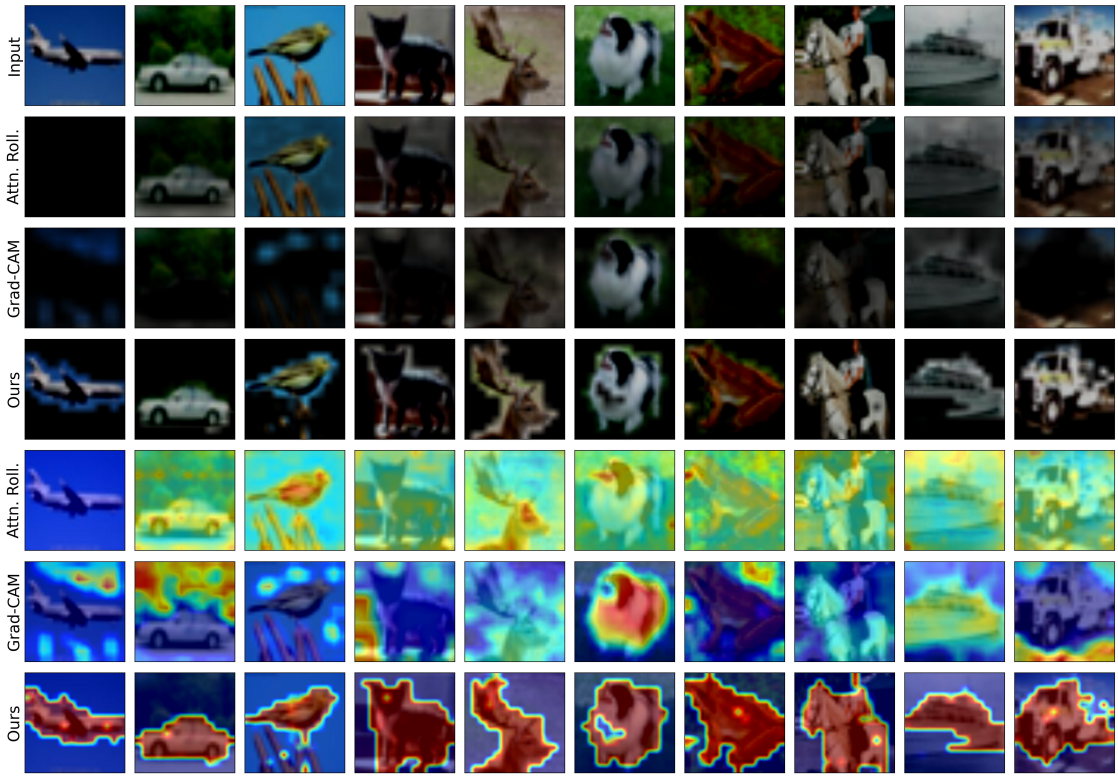

Figure 11: Empirical comparison of Vision DiffMask, Attention Rollout and Grad-CAM on CIFAR-10. The raw image is shown in the first row. Rows 2-4 correspond to the masked input, in the case of Vision DiffMask, or high-attribution patches for the other methods. Rows 5-7 overlay a heatmap over the original image based on attribution, with low scores depicted in blue and high in red.

Attribution methods cannot be evaluated simply by using human annotations, for that would measure the plausibility of the explanations according to humans, and not a faithful attribution according to the model Jacovi and Goldberg (2020). Hence, we first test the faithfulness of our model in a controlled scenario. Then, we evaluate our methodology using both qualitative and quantitative experiments against other state-of-the-art methods on CIFAR-10 (Krizhevsky, 2009). We show that our method produces plausible outputs and provides valuable insights into the model's decision-making procedure (see Figure 11 for a sample of the results).

This work is reported in (Nalmpantis et al., 2023).

3.1.2 A Joint Framework for Rationalization and Counterfactual Text Generation

In this work, we introduced CREST, a framework that merges selective rationalization with counterfactual text generation, enhancing the interpretability and robustness of NLP models. Our methodology centres around two core components:

1. **CREST-Generation**: We propose a novel method for generating counterfactual examples, which integrates sparse rationalization with span-level masked language modeling. This

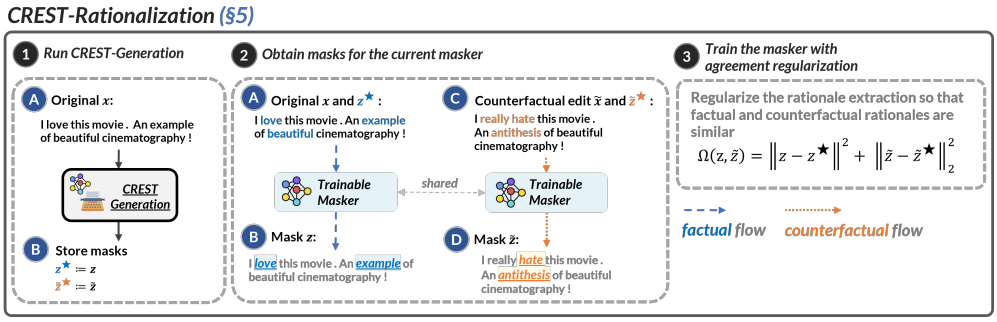

Figure 12: Overview of CREST-Rationalization.

approach yields counterfactuals that are not only valid and diverse but also maintain natural fluency, addressing the common challenges of counterfactual generation.

2. **CREST-Rationalization**: Building upon the generated counterfactuals, we introduce a regularization technique that decomposes rationalization into factual and counterfactual flows. This technique encourages consistency between rationales generated for both factual and counterfactual inputs, fostering more interpretable and robust model predictions.

We demonstrated that CREST efficiently produces high-quality, natural counterfactual examples and uses them to improve the quality and robustness of model rationales. Our findings indicate that models trained with CREST-generated counterfactuals show notable improvements in handling contrast and out-of-domain datasets. This work marks a significant step forward in combining two complementary approaches to achieve greater model transparency and reliability, offering new insights into the development of more interpretable and robust NLP systems.

We demonstrate the main components of the framework in Figure 12 and point the readers to our EMNLP 2023 paper (Treviso et al., 2023).

3.2 Transparent Evaluation

Automatic evaluation protocols are powered by trainable metrics, often built upon blackbox components such as pretrained LLMs. We contribute a novel paradigm for fine-grained evaluation of MT, powering fine-grained insights into model performance, an analysis of blackbox neural MT evaluation metrics, as well as a state-of-the-art approach for evaluation via fine-grained error detection. These contributions advance sub-goal 5.2b (transparent evaluation).

3.2.1 Extrinsic Evaluation of Machine Translation Metrics

Although machine translation (MT) is typically seen as a standalone application, in recent years MT models have been more frequently deployed as a component of a complex NLP platform delivering multilingual capabilities such as cross-lingual information retrieval or automated multilingual customer support. When an erroneous translation is generated by the MT systems, it may add new errors in the task pipeline leading to task failure and poor user experience. For example, consider the user's request in Chinese 剑桥有牙买加菜吗？ (*"Is there any good Jamaican food in Cambridge?"*) machine-translated into English as *"Does Cambridge have a good meal in Jamaica?"*. The model will erroneously consider "Jamaica" as a location, instead of cuisine, and

prompt the search engine to look up restaurants in Jamaica. To avoid this *breakdown*, it is crucial to detect an incorrect translation before it causes further errors in the task pipeline.

One way to approach this *breakdown detection* is using segment-level scores provided by MT metrics. Recent MT metrics have demonstrated high correlation with human judgements at the system level for some language pairs (Ma et al., 2019). These metrics are potentially capable of identifying subtle differences between MT systems that emerge over a relatively large test corpus. These metrics are also evaluated on respective correlation with human judgements at the segment level, however, there is a considerable performance penalty (Ma et al., 2019; Freitag et al., 2021). Segment-level evaluation of MT is indeed more difficult and even humans have low inter-annotator agreement on this task (Popović, 2021). Despite MT systems being a crucial intermediate step in several applications, characterising the behaviour of these metrics under task-oriented evaluation has not been explored. In this work, we provide a complementary evaluation of MT metrics. We focus on the segment-level performance of metrics, and we evaluate their performance extrinsically, by correlating each with the outcome of downstream tasks with respective, reliable accuracy metrics. We assume access to a parallel task-oriented dataset, a task-specific monolingual model, and a translation model that can translate from the target language into the language of the monolingual model. We consider the *Translate-Test* setting — where at test time, the examples from the test language are translated to the task language for evaluation. We use the outcomes of this extrinsic task to construct a breakdown detection benchmark for the metrics.

We use dialogue state tracking, semantic parsing, and extractive question answering as our extrinsic tasks. We evaluate nine metrics consisting of string overlap metrics, embedding-based metrics, and metrics trained using scores from human evaluation of MT. Surprisingly, we find our setup challenging for all existing metrics; demonstrating poor capability in discerning good and bad translations across tasks. We present a comprehensive analysis of the failure of the metrics through quantitative and qualitative evaluation.

Our contributions are summarised as follows:

1. We derive a new breakdown detection task, for evaluating MT metrics, measuring how indicative segment-level scores are for downstream performance of an extrinsic cross-lingual task. We evaluate nine metrics on three extrinsic tasks covering 39 unique language pairs.

2. We show that segment-level scores, from these metrics, have minimal correlation with extrinsic task performance. Our results indicate that these scores are uninformative at the segment level — clearly demonstrating a serious deficiency in the best contemporary MT metrics. In addition, we find variable task sensitivity to different MT errors .

3. We propose recommendations on developing MT metrics to produce useful segment-level output by predicting labels instead of scores and suggest reusing existing post-editing datasets and explicit error annotations.

See Moghe et al. (2023) for further details.

3.2.2 The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics

In this work, we investigate neural metrics for machine translation (MT) evaluation, such as COMET, which have demonstrated superior correlation with human judgments over traditional

metrics like BLEU. Despite their advancements, a considerable challenge remains with these neural metrics due to their "black box" nature, offering minimal transparency into their decision-making processes. Our research aims to develop and compare various neural explainability methods to shed light on the interpretive capabilities of state-of-the-art fine-tuned neural metrics, particularly focusing on how these metrics utilize token-level information to compute sentence-level scores.

We evaluated two prominent neural metrics, COMET and UNITE, using a suite of attribution and input attribution methods to generate token-level explanations. These explanations were then analyzed in conjunction with human-annotated error spans from Multidimensional Quality Metrics (MQM) annotations and synthetically generated critical translation errors. Our methodology is grounded in the hypothesis that the neural metrics leverage token-level information, which can be unveiled through our explainability techniques, such as:

- **Embedding alignments**: We measure the maximum cosine similarity between the embeddings of each translation token and the reference and/or source tokens (**embed-align**).
- **Gradient-based**: We calculate the ℓ_2-norm of gradients with respect to the word embeddings of the translation tokens (**grad**ℓ_2), aiming to identify the tokens that have the most significant impact on the final score.
- **Attention-based**: These include the use of attention weights (**attn**) and attention weights scaled by the ℓ_2-norm of the translation tokens (**attn**x**grad**), these techniques aim to highlight the importance of specific tokens based on the model's attention mechanism.

Our findings reveal a close relationship between the quality of explanations and the metric's architecture. Notably, incorporating reference information significantly improves the quality of explanations. Furthermore, our token-level explanations align with human-annotated error spans, effectively identifying critical translation errors and highlighting potential weaknesses within the metrics.

Our analysis underscores the impact of reference information in enhancing the explainability of neural MT metrics. The experiments show that explanations for critical translation errors, such as negations and hallucinations, can reveal potential shortcomings in these metrics. Specifically, we find that neural metrics are adept at identifying and penalizing hallucinated translations, aligning with the observed phenomena in critical error detection tasks.

For more information we direct readers to (Rei et al., 2023).

3.2.3 xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection

Following the previous insights regarding the quality predictions of COMET (see Section 3.2.2), we further explored machine translation (MT) evaluation by integrating the strengths of traditional sentence-level evaluation metrics and the granular error detection capabilities of generative large language models (LLMs). We introduce xCOMET, a novel metric that is able to provide quality estimates at sentence and word level that correlate well with human judgements while detecting error spans within translations.

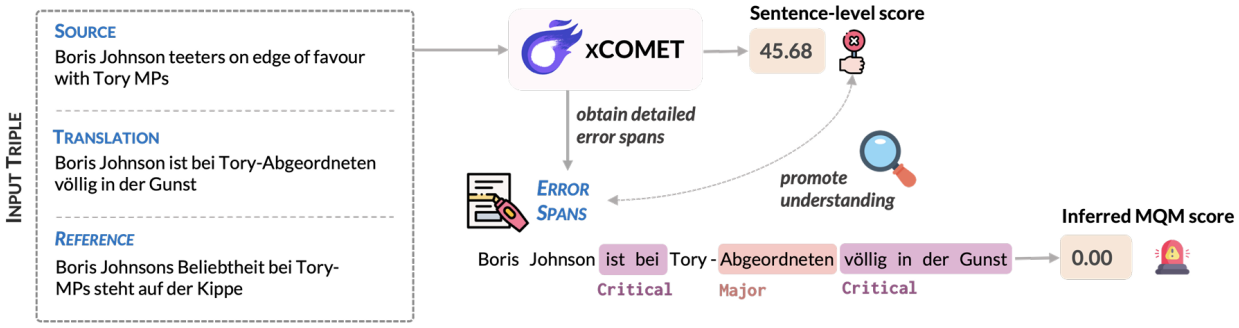

Figure 13: The xCOMET framework illustrated through a real example: the metric not only provides a sentence-level score but also predicts translation error spans along with their respective severity. From these spans, we can infer MQMscore (following the MQM typology) that informs and highly correlates with the sentence-level score. These spans complement the sentence-level score by providing a detailed view into the translation errors.

Our proposed metric demonstrates high performance across several evaluation dimensions including sentence-level, system-level, and error span prediction, setting a new standard in the field of MT evaluation (this model won the competitive WMT23 Metrics Shared Task (Freitag et al., 2023) beating submissions by Google and Microsoft). We further ensure its robustness to hallucinations and detection of critical errors. In an effort to contribute to the broader research community and foster further development in this area, we have publicly released two models of xCOMET, namely xCOMET-XL and xCOMET-XXL.

The advent of xCOMET marks a significant advancement for context-aware machine translation by offering several key benefits. Firstly, the metric's capability to highlight and categorize error spans equips developers and researchers with the means to precisely pinpoint and address specific weaknesses in translation systems, facilitating targeted improvements. Secondly, xCOMET's demonstrated robustness in accurately identifying a diverse array of errors, including the less commonly addressed hallucinations, ensures that evaluations provide a reliable measure of translation quality across various scenarios. Lastly, by making our models available to the public, we aim to encourage the adoption and iterative enhancement of xCOMET, promoting collaborative efforts within the community to refine and advance MT evaluation methodologies.

We illustrate the xCOMET framework in Figure 13.

See Guerreiro et al. (2023d) for further details.

Plans for future work

This task remains important for UTTER's goals and we plan to continue working on it in the second half of the project. We expect more synergy between this task and T5.1, since interpretable uncertainty quantification may well benefit from tools for model interpretation (*e.g.*, input attribution).

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 projet till month 24. Indeed, the first half of the project saw considerable progress along the two dimensions above (*e.g.*, hallucinations and robustness).

Next, we report on our progress in the first 18 months of the project. Our contributions are organised under 2 themes, one focused on 5.3a and one on 5.3b (with some natural overlap).

4.1 Hallucinations

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 the next three pieces we contribute a comprehensive study of hallucinations, a categorisation of different types of hallucinations, as well as methodology and software for detection and mitigation. These contributions advance sub-goal 5.3a (hallucinations).

4.1.1 A Comprehensive Study of Hallucinations in Neural Machine Translation

In this work, we set the foundation for the more systematic study of hallucinations in neural machine translation, addressing the challenge of detecting and understanding hallucinatory outputs in translations. Our approach is distinguished by its focus on natural, in-domain data, free from artificial perturbations either during training or inference. We annotate a dataset of over 3.4K sentences, identifying different kinds of critical errors and hallucinations, and evaluate detection methods, revisiting previously used methods and proposing the use of glass-box, uncertainty-based detectors.

For hallucination detection, we cover previously proposed heuristic methods and also introduce simple model uncertainty measures as detectors. Our analysis reveals that in preventive settings, many previously used methods are largely inadequate, while standard sequence log-probability emerges as the most effective, performing on par with reference-based methods. his approach provided us with a multi-faceted view of the hallucination phenomenon in NMT systems (see also Figure 14).

Our analytical efforts yielded several insights:

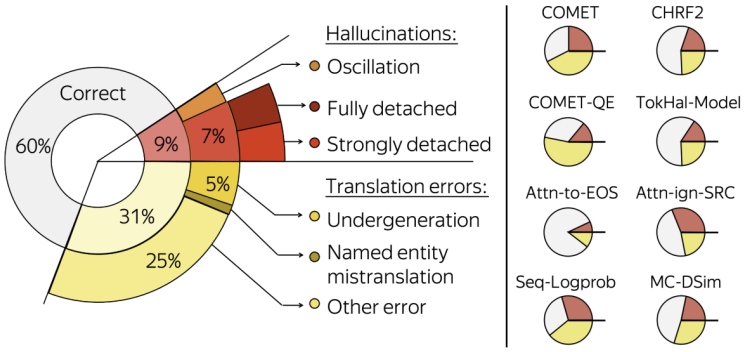

Figure 14: Overall (left) and method-specific (right) statistics of human annotation results. Method-specific statistics show the percentages of correct translations (grey), translation errors (yellow) and hallucinations (red) among the examples flagged by each method.

- The prevalence of hallucinations in unperturbed, natural settings is substantiated, challenging the necessity of artificial amplification used in earlier studies. This finding underscores the inherent vulnerability of NMT systems to this pathological output.
- In the quest for effective hallucination detection, our evaluation revealed that previously used methods fell short in accuracy and reliability. Surprisingly, sequence log-probability, a relatively straightforward measure, emerged as the most effective detector, rivalling even reference-based methods in performance.
- We introduced *DEHALLUCINATOR*, an innovative method designed to mitigate hallucinations at test-time. This approach leverages a lightweight hallucination detector, followed by a mechanism to replace flagged translations with more accurate alternatives. Our experiments demonstrated a significant reduction in the rate of hallucinations, enhancing the overall quality and reliability of NMT outputs.

This analysis not only advances our understanding of hallucinations in NMT but also sets a robust foundation for future research in this area of study. More information is available at Guerreiro et al. (2023e).

4.1.2 Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation

In this work, we tackled the problem of hallucination detection in neural machine translation from a novel perspective. We argue that since hallucinations significantly diverge from the source content, they manifest distinct cross-attention patterns compared to non-hallucinatory, high-quality translations. Leveraging the theoretical foundation of Optimal Transport (OT), we propose a fully unsupervised, plug-in detector applicable to any attention-based NMT model, aiming to identify translations with aberrant attention mass distributions.

We define the following scenarios for analysis:

1. **Wass-to-Unif**: Compares the source attention mass distribution of a given translation to a uniform distribution, under the premise that hallucinatory outputs tend to exhibit abnormal concentration on irrelevant tokens.

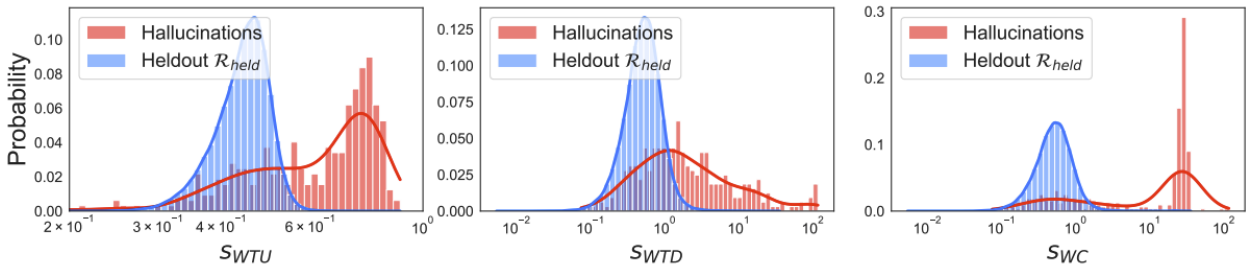

Figure 15: Histogram scores for our methods – Wass-to-Unif (left), Wass-to-Data (center) and Wass-Combo (right). We display Wass-to-Data and Wass-Combo scores on log-scale.

2. **Wass-to-Data**: Uses a data-driven reference distribution, comparing the source attention mass distribution of a translation to those derived from a corpus of high-quality translations. This method aims to contextualize the anomaly within a model's typical distribution of attention patterns.

3. **Wass-Combo**: We combine the two scores in two steps: (i) we start by assessing whether a test sample is deemed a hallucination according to Wass-to-Unif, and if not (ii) we compute the Wass-to-Data score

Our experimental evaluation, conducted across various datasets and language pairs, is summarised in Figure 15 demonstrates that:

- The proposed OT-inspired hallucination detector significantly outperforms existing model-based detectors in identifying hallucinatory translations.
- The detector proves competitive with existing external detectors that require auxiliary models trained on extensive datasets for related tasks, such as quality estimation and cross-lingual sentence similarity.
- Our findings affirm the hypothesis that hallucinatory translations exhibit cross-attention patterns that are statistically distinct from those of high-quality translations, validating the effectiveness of our OT-based anomaly detection approach.

Our study introduces a pioneering approach to hallucination detection in NMT, showcasing the utility of optimal transport theory in analyzing the complex phenomena of hallucinations. By establishing a method that does not rely on supervised learning or extensive auxiliary datasets, we offer a scalable, efficient solution that enhances the robustness and trustworthiness of NMT systems. This work paves the way for further exploration into unsupervised methods for ensuring translation quality and model reliability.

This work is reported in (Guerreiro et al., 2023c).

4.1.3 Hallucinations in Large Multilingual Translation Models

In this work, we turn to large language models, motivated by the significant gap in understanding how hallucinations manifest in multilingual models and their implications on translation quality, especially beyond English-centric language pairs and in low-resource contexts. We specifically

analyze hallucinations for both M2M neural machine translation (NMT) models and generative pre-trained transformers (GPTs) across over 100 language pairs.

Our methodology encompasses a comprehensive evaluation framework that categorizes hallucinations into two distinct types: *hallucinations under perturbation* and *natural hallucinations*. We employ a range of artificial perturbations (e.g., misspellings, insertion of tokens, and capitalization errors) to assess the robustness of translation models under manipulated conditions. Additionally, we investigate natural hallucinations by examining the models' output on unperturbed source texts, focusing on their propensity to generate content unrelated to the source.

Our analysis employs several metrics, including spBLEU Goyal et al. (2022) for lexical similarity and COMET variants for semantic evaluation Rei et al. (2022a,b), alongside sentence similarity scores computed by LaBSE Feng et al. (2022).

Our findings reveal several insights into the behaviour of large multilingual MT models regarding hallucinations:

- Hallucinations are more prevalent in low-resource language pairs and when translating out of English, underscoring the challenge of maintaining translation quality in less-represented languages.
- Distinctly, LLMs such as GPT exhibit qualitatively different hallucinations compared to conventional NMT models, including off-target translations and overgeneration, suggesting differing underlying mechanisms of error generation.
- Smaller distilled models like SMaLL100 demonstrate lower rates of hallucinations than their larger counterparts, hinting at the potential benefits of model distillation in mitigating translation pathologies.
- Employing diverse models trained on different data or with varied procedures as fallback systems can significantly improve translation quality and effectively mitigate certain pathological outputs, demonstrating a promising approach for enhancing the robustness of MT systems.

Our comprehensive analysis across diverse linguistic scenarios and model types sheds light on hallucinations in machine translation. It highlights the critical need for robust mechanisms to detect and mitigate hallucinations, especially in low-resource contexts beyond English-centric cases, to ensure the reliability and safety of MT systems in real-world applications. More information is available at Guerreiro et al. (2023a)

4.2 Robustness

In the next two pieces, we explore variants of system combination to improve robustness to noise and out-of-domain inputs in decoding and in evaluation. These contributions advance sub-goal 5.2b (robustness).

4.2.1 Translation Hypothesis Ensembling with Large Language Models

Large language models (LLMs) are becoming a one-fits-many solution, but they sometimes hallucinate or produce unreliable output (Guerreiro et al., 2023b). In this paper, we focus on the specific

task of machine translation and investigate how hypothesis ensembling can improve the quality of the generated text. We experiment with several techniques for ensembling hypotheses produced by LLMs such as ChatGPT[3], LLaMA (Touvron et al., 2023), and Alpaca (Taori et al., 2023), providing a comprehensive study along multiple dimensions: the method to generate hypotheses (multiple prompts, temperature-based sampling, and beam search) and the strategy to produce the final translation (instruction-based, quality-based reranking (Fernandes et al., 2022), and minimum Bayes risk (MBR) decoding (Eikema and Aziz, 2022)).

Our main findings can be summarized as follows. First, we demonstrate that translation quality can be enhanced with a small number of samples (*e.g.*, 20), especially when translating out of English. Notably, this differs from the findings of previous research using task-specific NMT models (Fernandes et al., 2022; Freitag et al., 2022). Second, we discuss in which conditions beam search remains a reliable baseline for single-hypothesis translation and how to ensemble translations. Moreover, we find that there exists a significant gap in the quality of ensembles of unbiased samples from LLaMA and Alpaca. We attribute this disparity to how instruction tuning affects the relationship between the diversity of the hypotheses and the sampling temperature, which ultimately impacts translation quality. Lastly, we show that hypothesis ensembling reduces the number of generated hallucinations, thereby improving the model's robustness to source perturbations. Ensembling predictions and increasing the model size narrows the quality gap between open-source models and ChatGPT.

A full description of this work can be found in our paper (Farinhas et al., 2023).

4.2.2 Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation

In this work, we explored methods to enhance the reliability of state-of-the-art machine translation evaluation metrics. While neural-based metrics like COMET Rei et al. (2020) and BLEURTSellam et al. (2020) have shown strong correlations with human judgments, they struggle to detect critical errors, such as deviations in entities and numbers. In contrast, traditional metrics like BLEU and chrF, despite their lower correlations with human judgments, exhibit sensitivity to these deviations. To address this, we examined various strategies for integrating both approaches. By incorporating additional information, such as sentence-level features and word-level tags, during training, we found that the metrics improved in penalizing translations with erroneous patterns of interest.

To be more specific, we considered a range of approaches leveraging interpretable string-based metrics to fortify the robustness of contemporary neural-based metrics like COMET. These approaches include ensembling metrics, integrating sentence-level features, and leveraging word-level information from TER-based alignmentsSnover et al. (2006). Notably, we discovered that slight modifications to the COMET architecture, such as incorporating sentence-level features based on BLEU and chrF scores or integrating word-level tags for the hypothesis, yielded competitive performance enhancements. To validate the efficacy of our proposed methods, we conducted evaluations on the latest MQM test set, covering diverse domains and language pairs, as well as on challenge sets from the WMT 2022 Metrics shared task . Using these approaches resulted in notable gains in correlation with human judgments, as well as promising performance on the aforementioned challenge sets across multiple language pairs. The results were encouraging, underscoring the potential of our approaches in advancing machine translation evaluation metrics.

[3] https://chat.openai.com/

Extending this work we are currently studying further the impact of using word-level tags of the hypothesis in other ways not covered in this paper, e.g., by encoding this supplementary data as word factors Niehues et al. (2016). The variety of the word-level tags potentially could also be extended beyond binary quality tags to cover specific entities in text such as numbers and named entities.

This work is reported in (Glushkova et al., 2023).

Plans for future work

This task remains very relevant for UTTER's goals and we expect to continue working on it in the second half of the project as originally planned. For the second half of the project, we plan to focus more on speech.

5 Conclusion

WP5 saw progress on all three tasks, we have contributed datasets, methodology, software and empirical observations to advance various aspects of uncertainty-aware generation, explainability and robustness. The first half of the project was focused on the text modality, hence we expect to transfer some of this progress to speech in the second half. There are no relevant risks to be listed for WP5, and we expect steady progress in the second half of the project.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889, 12 2021. ISSN 0891-2017. doi: 10.1162/coli_a_00418. URL https://doi.org/10.1162/coli_a_00418.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

John Langshaw Austin. *How to do things with words*. Clarendon Press, Oxford, 1975.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.124.

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.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero, editors, *Artificial Neural Networks and Machine Learning – ICANN 2016*, pages 63–71, Cham, 2016. Springer International Publishing.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1860–1874. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-long.146. URL https://doi.org/10.18653/v1/2021.acl-long.146.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics, 2021b. doi: 10.18653/V1/2021.EMNLP-MAIN.522. URL https://doi.org/10.18653/v1/2021.emnlp-main.522.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang,

and Xing Xie. A survey on evaluation of large language models. *CoRR*, abs/2307.03109, 2023. doi: 10.48550/arXiv.2307.03109. URL https://doi.org/10.48550/arXiv.2307.03109.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022a. doi: 10.18653/V1/2022.ACL-LONG.581. URL https://doi.org/10.18653/v1/2022.acl-long.581.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022b. doi: 10.18653/V1/2022.ACL-LONG.581. URL https://doi.org/10.18653/v1/2022.acl-long.581.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. 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 36–50, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.3. URL https://aclanthology.org/2023.acl-long.3.

Soham Dan and Dan Roth. On the effects of transformer size on in- and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.180. URL https://aclanthology.org/2021.findings-emnlp.180.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL https://aclanthology.org/2020.emnlp-main.262.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112, 2009.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21.

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. *ICLR*, 2021.

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 and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.754.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/632.html.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation, October 2020. URL http://arxiv.org/abs/2010.11125. arXiv:2010.11125 [cs].

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, 2022.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.100. URL https://aclanthology.org/2022.naacl-main.100.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.73.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022. doi: 10.1162/tacl_a_00491. URL https://aclanthology.org/2022.tacl-1.47.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

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.

Yoav Goldberg and Graeme Hirst. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 2017. ISBN 1627052984.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

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, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023b.

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 2023c. 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*, 2023d.

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 2023e. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.75. URL https://aclanthology.org/2023.eacl-main.75.

Joseph Y Halpern. *Reasoning about uncertainty*. MIT press, 2017.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 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 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.22.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Moksh Jain, Salem Lahlou, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, December 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Kristiina Jokinen. Goal formulation based on communicative principles. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996. URL https://aclanthology.org/C96-2101.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/4d13b2d99519c5415661dad44ab7edcd-Abstract-Conference.html.

Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7811–7818. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.698. URL https://doi.org/10.18653/v1/2020.acl-main.698.

Andrey N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Pub Co, 2 edition, June 1960.

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

Willem JM Levelt. *Speaking: From intention to articulation*. MIT press, 1993.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia, editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics, 2017. doi: 10.18653/V1/K17-1034. URL https://doi.org/10.18653/v1/K17-1034.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive Decoding: Open-ended Text Generation as Optimization, October 2022. URL http://arxiv.org/abs/2210.15097. arXiv:2210.15097 [cs].

Dennis V Lindley. *Understanding uncertainty*. John Wiley & Sons, 2013.

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

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302. URL https://aclanthology.org/W19-5302.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. On the probability–quality paradox in language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.5. URL https://aclanthology.org/2022.acl-short.5.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

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.

Jan Niehues, Thanh-Le Ha, Eunah Cho, and Alex Waibel. Using factored word representation in neural network language models. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 74–82, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2208. URL https://aclanthology.org/W16-2208.

OpenAI. Introducing chatgpt. *Available at https://openai.com/blog/chatgpt*, 2022. URL https://openai.com/blog/chatgpt.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250. URL https://doi.org/10.18653/v1/D19-1250.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.731.

Maja Popović. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.18. URL https://aclanthology.org/2021.conll-1.18.

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.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal nucleus sampling. *arXiv preprint arXiv:2305.02633*, 2023.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 2022a.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, 2022b.

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.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/412243.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

John R Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, 1969.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*, 2014.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

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.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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.

Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.326. URL https://aclanthology.org/2020.acl-main.326.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. Cross-lingual knowledge editing in large language models. *CoRR*, abs/2309.08952, 2023a. doi: 10.48550/ARXIV.2309.08952. URL https://doi.org/10.48550/arXiv.2309.08952.

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

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3082–3095. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/148c0aeea1c5da82f4fa86a09d4190da-Paper-Conference.pdf.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3: 283–297, 2015. doi: 10.1162/tacl_a_00139. URL https://aclanthology.org/Q15-1021.

Davis Yoshida, Kartik Goyal, and Kevin Gimpel. Map's not dead yet: Uncovering true language model modes by conditioning away degeneracy, 2023.

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.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *CoRR*, abs/2305.12740, 2023. doi: 10.48550/ARXIV.2305.12740. URL https://doi.org/10.48550/arXiv.2305.12740.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. Modifying memories in transformer models. *CoRR*, abs/2012.00363, 2020. URL https://arxiv.org/abs/2012.00363.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D19 First Report on Uncertainty-Aware, Robust and Explainable Models