Beyond statistical significance: Quantifying uncertainty and statistical variability in multilingual and multitask NLP evaluation

Jonne Sälevä
Brandeis University
jonnesaleva@brandeis.edu

Duygu Ataman
Middle East Technical University
dataman@metu.edu.tr

Constantine Lignos
Brandeis University
lignos@brandeis.edu

Abstract

We introduce a set of resampling-based methods for quantifying uncertainty and statistical precision of evaluation metrics in multilingual and/or multitask NLP benchmarks. We show how experimental variation in performance scores arises from both model and data-related sources, and that accounting for both of them is necessary to avoid substantially underestimating the overall variability over hypothetical replications. Using multilingual question answering, machine translation, and named entity recognition as example tasks, we also demonstrate how resampling methods are useful for quantifying the replication uncertainty of various quantities used in leaderboards such as model rankings and pairwise differences between models.

1 Introduction

Over the last several years, multilingual research has undergone exponential growth within NLP, both in terms of model capabilities as well as evaluation datasets. Since no paper can evaluate on all languages, it is important to determine to what extent research findings generalize from one language and from one task to another. This is particularly true in recent years with the proliferation of large language models (LLMs), where the goal often is to draw conclusions about the models that are not task or language-dependent.

Evaluation paradigms Despite this need, evaluation setups in multilingual NLP research tend to focus on within-language evaluation and fall into two broad categories. In the first category, the same experiments are repeated across several languages and the results analyzed separately. The inferences are typically reported as one collection, e.g. "Model A significantly outperforms model B on English–Finnish and English–Turkish but not English–German" or "Model A outperforms model B on 13 out of 43 language pairs." The second variant, depicted in Figure 1, is the "leaderboard-type" scenario in which performance is evaluated on each of the L languages after which the numbers are distilled into a single performance measure using a pre-specified aggregation function such as the arithmetic mean. Given the aggregate scores, the main interest may then be to *rank* the M models based on the single score or to compare differences between observed scores.

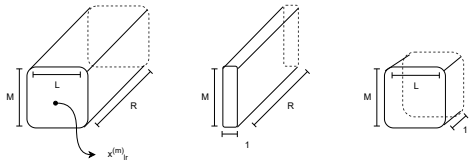

Figure 1: Left: Notional diagram of a $M \times L$ leaderboard consisting of M models tested on L languages and replicated R times each, yielding individual observations $x_{lr}^{(m)}$. Middle: Aggregation over L languages into a single scalar per model and estimating between-language variance ν_m^2. Right: Aggregation over replications to estimate within-language uncertainty η_{ml}.

Statistical significance and effect sizes The most interesting questions to practitioners often revolve around whether a given model "truly" outperforms another model or whether an observed performance difference was simply "random noise" that could be expected to occur under replication when comparing two equally performing models. From a statistical perspective, estimating this "consistency with random noise" is intimately linked to testing for statistical significance. Under conventional null hypothesis significance testing the typical "$p \leq 0.05$" threshold is achieved when an estimate $\hat{\theta}$ lies at least 1.96 (≈ 2) standard errors from a hypothesized null value θ_0, often zero. Following Gelman (2018) we refer to the ratio $|\hat{\theta}/\text{se}(\hat{\theta})|$ as *effect size* and use it to judge statistical significance based on whether it exceeds a threshold of 2.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2304–2321

To estimate such effect sizes requires specifying the source(s) of random variation in our experiments. Typically, this variation has been treated as arising from one of two sources: *data-side variability* and *model-side variability* which we describe below.

Data-side variability The former quantifies how differently each model would perform if a slightly different test set were used for a given task. In practice this is typically done using resampled versions of the original test set, for example using the *bootstrap* resampling algorithm (Efron, 1979).

From a sampling perspective, data-side variability can be interpreted as sampling error associated with the process of constructing a test set for a given task by sampling data from a larger set of all possible test sets. Another source of data-related variability that appears in leaderboards and other multi-task evaluations is the choice of evaluation tasks. When ranking is performed based on aggregated quantities, the overall performance and rankings of models will vary as a function of which tasks are chosen. This can be understood as the sampling error associated with creating an evaluation benchmark out of the larger population of all possible ones that could have been constructed.

Model-side variability There is also *model-side variability* that will be present even if the data set is held constant. With LLMs, decoding often involves random sampling of tokens, especially with generative tasks like question answering. By drawing multiple responses for each question, the performance will fluctuate randomly around some average. Even if the inference algorithm is deterministic (e.g. beam search), any potential finetuning will likely be nondeterministic due to, for example, random shuffling of minibatches and weight initialization. This will yield slightly different parameters from run to run, which will potentially yield different responses and different values of the performance metric. From a sampling perspective, model-side uncertainty can be seen as drawing samples from the probability distribution over possible responses defined by the model.

Summary In this paper, we provide a new perspective on how resampling-based methods can be useful in analysis of experimental results in multilingual and multitask NLP.

Through connections to statistics, we show how experimental variance can be decomposed into between and within-task components (ν and η_l) and how the latter further decomposes into model and data-based components (σ_l, τ_l) which arise naturally from the structure of the experimental data.

Using question answering (QA), machine translation (MT), and named entity recognition (NER) as case studies, our results demonstrate that none of these sources of variation are negligible. This suggests that only analyzing one source of error may underestimate the total variation and expose researchers to the risk of drawing incorrect conclusions about the relative merits of models.

We show how resampling and estimates of between and within-language variance components can be used to derive uncertainty-aware estimates of more complex quantities such as relative rankings of models and approximate distributions of pairwise performance differences between them.

All of our methods run in seconds to minutes and present no major computational bottlenecks on top of the overall inference time complexity. We provide an implementation of our approach in a toolkit which we make freely available at `https://github.com/j0ma/reuben`.[1]

2 Related work

Significance testing in ML and NLP Within ML, statistical significance and hypothesis testing have been prevalent since at least the 1990s (e.g. Dietterich, 1998; Dror et al., 2018; Demšar, 2006), though the emphasis has often been on "finding the right test" instead of fully modeling the available data. Within NLP, significance testing is often done using permutation tests and bootstrap-based hypothesis tests (Noreen, 1989; Koehn, 2004), which have become the de-facto standard featured in state-of-the-art evaluation toolkits (e.g. Post, 2018). Significance testing based on resampling is also often used when evaluating task-specific MT metrics against human judgments in order to gauge what difference magnitudes correspond to real differences as perceived by humans(e.g. Lo et al., 2023; Kocmi et al., 2021).

In contrast, Ulmer et al. (2022) advocate for understanding model performance variation based on multiple model training runs instead of resampling, and using this model-side variation to assess statistical significance. There is no clear consensus though, and others (e.g. Bethard, 2022) expressly

[1]The name `reuben` is an acronym for "**RE**sampling-based **U**ncertainty **B**ounds for **E**valuating **N**LP.

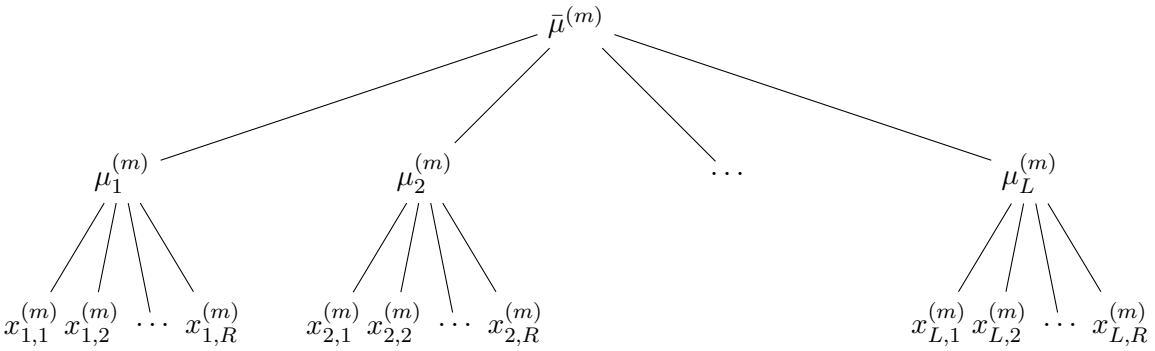

Figure 2: Tree diagram showing the multilevel structure of the experimental data containing R replications of L languages for a single model. Between-language variance ν_m^2 is computed across the averages $\mu_l^{(m)}$ whereas within-language variance η_{ml}^2 corresponds to the variance among the leaf nodes $x_{lr}^{(m)}$ of each subtree.

warn against using random seeds to provide estimates of score distributions based on the argument that random seeds should be optimized like other hyperparameters. Some work in NLP has also evaluated different sources of variation in evaluation of MT metrics (Xiang et al., 2022).

Evaluation practices in NLP There has also been a healthy debate regarding best practices in test set construction and more generally of evaluation using held-out datasets. Gorman and Bedrick (2019) and others (e.g. Kodner et al., 2023; Liu and Dorr, 2024) argue that we should *randomize* our splits and evaluate on several versions of our test sets to reduce the chance of false positives. Søgaard et al. (2021) instead favor using approaches based on adversarial splitting, arguing that naive randomization will underestimate the variation.

Benchmarks have received much attention in the general machine learning community. An excellent survey is Dehghani et al. (2021) where the authors discuss the lifecycle of overall ML benchmarks, how they become stale as well as what is random when models are evaluated. The question of what aggregation function to use in leaderboards is also explored by Tatiana and Valentin (2021) where the authors explore using both the arithmetic, geometric and harmonic means. In particular, they show that results may be wildly different depending on what aggregation measure is used. Longjohn et al. (2025) also survey and develop resampling-based and Bayesian methods for popular computer vision benchmarks for LLMs.

Related ideas in statistics While the original bootstrap algorithm (Efron, 1979) was intended as a frequentist estimation device, a Bayesian interpretation was provided by Rubin (1981), allowing the resulting estimates (empirical distributions, intervals etc.) to be interpreted as posterior quantities. Uncertainty quantification (i.e. standard error estimation) using bootstrap and other nonparametric methods is thoroughly reviewed by Efron (1981). On the importance of understanding variation, Gelman (2005) provides an excellent argument for the importance of understanding variance components in statistical analysis more generally. For a more specific application to linguistic research, see Vasishth and Gelman (2021). The multilevel structure we use is also related to the technique of *random-effects meta-analysis*, see Higgins et al. (2008).

Uncertainty quantification in NLP There exists a substantial literature (e.g. Nikitin et al., 2024; Da et al., 2025; Chen et al., 2024, 2025; Yang et al., 2025; Ye et al., 2024; Wagner et al., 2024; Blackwell et al., 2025) on a task referred to as *uncertainty quantification* in NLP, where the task is to estimate some properties of an LLM's distribution over continuations, $p(x|x_{\text{prompt}})$, that predicts whether the model is able to produce the correct answer to this question. While this task quantifies per-example uncertainty, our efforts instead focus on *experimental variation*. i.e. estimating the noise we would expect to see under repeated experiments and hypothetical evaluation suites.

3 Statistical background

We begin by formalizing the analysis of multilingual experimental results as a statistical inference problem related to a hierarchically structured population of experimental results.

Observations Our experimental data consists of scores generated by separately evaluating M models on L languages. Individual model-level observations are represented as L-vectors of scores:

$$\mathbf{x}^{(m)} = (x_1^{(m)}, \ldots, x_L^{(m)})^\top \in \mathbb{R}^L$$

The entire table of experimental results is then an M-by-L matrix where each row corresponds to a model $\mathbf{x}^{(m)}$, as shown in Figure 1. Hierarchically, our data can be seen as arising from a three-level population $\bar{\mu}^{(m)} \mapsto \mu_l^{(m)} \mapsto x_{lr}^{(m)}$. This structure is depicted in Figure 2 in the Appendix. At the topmost level, each model is associated with a population-level average performance ($\bar{\mu}^{(m)}$), and individual languages are seen as subsets of this larger population, each with their own averages ($\mu_l^{(m)}$). At the lowest level, we have the observed scores ($x_{lr}^{(m)}$) which represent replications of an experiment on a given subpopulation. Leaderboard-style evaluation with aggregate performance measures can then be seen as estimating $\bar{\mu}^{(m)}$ by using an aggregation function as the estimator.

Sources of uncertainty Under infinite replications, our observed values $x_l^{(m)}$ values will center around some mean value and exhibit a degree of fluctuation around it, representing measurement error arising from both model-side randomness (e.g. nondeterministic decoding) as well as sampling error of the test set. We can represent this as $x_{lr}^{(m)} = \mu_l^{(m)} + \varepsilon_{lr}^{\text{repl}}$ where $\varepsilon_{lr}^{\text{repl}}$ represents the deviation of observed scores from the language-specific mean $\mu_l^{(m)}$. In addition to this "within-language" variation arising from replication, there is also "between-language" variability related to what tasks are included in the leaderboard. Concretely, each $\mu_l^{(m)}$ also differs from the global mean performance across all tasks by some amount, i.e. $\mu_l^{(m)} = \bar{\mu}^{(m)} + \varepsilon_l^{\text{lang}}$. This lets us decompose the overall observations into $x_{lr}^{(m)} = \bar{\mu}^{(m)} + \varepsilon_l^{\text{lang}} + \varepsilon_{lr}^{\text{repl}}$. Taking the variance of both sides, we obtain a first-principles variance decomposition into between and within-language components $\mathbb{V}[x_{lr}^{(m)}] = \mathbb{V}[\varepsilon_l^{\text{lang}}] + \mathbb{V}[\varepsilon_{lr}^{\text{repl}}] = \nu_m^2 + \eta_{lm}^2$ where ν_m^2 and η_{lm}^2 refer to between and within-language variation of model m, respectively.

Replication For a given task, two orthogonal sources of replication noise are apparent. *Model-side randomness* that arises from nondeterminism in decoding/text generation. Additional model-side randomness may arise from random parts of any training/finetuning runs, related to randomized weight initialization and batching. While less common in inference-only LLM evaluation, this is still a relevant source of variability for older non-pretrained models such as LSTMs (e.g. Reimers and Gurevych, 2017) and older pretrained models such as BERT and XLM-R. Another common source of score variability is *sampling error* when constructing each test set $\mathcal{D}_l$ (e.g. Koehn, 2004; Dehghani et al., 2021). Since we tend not to have access to true sampling distribution $p(\mathcal{D}_l)$, this variability is typically estimated by *resampling* with replacement from the original data and computing an empirical variance estimate. This is the approach taken by many evaluation libraries such as `sacrebleu` (Post, 2018) and `lm-evaluation-harness` (Gao et al., 2024).

From a sampling perspective, these sources of variability can be viewed as parts of the data collection process: first, given a model m and a random seed s, we sample random responses to all examples in our test set, i.e. $x_{ls} \sim p(x_l)$. Finetuning can be seen as sampling $\theta_{ls}^{(m)} \sim p(\theta_l^{(m)})$ from the distribution of all model parameters that could be obtained. On the dataset side, we sample a set of languages to evaluate on and, for each language, a test set from a larger hypothetical population of similar data in language l, $\mathcal{D}_{lb} \sim p(\mathcal{D}_l)$. This decomposes the total variance further as

$$\mathbb{V}[x_{lr}^{(m)}] = \underbrace{\nu_m^2}_{\text{lang}} + \underbrace{\sigma_{ml}^2}_{\text{seed}} + \underbrace{\tau_{ml}^2}_{\text{boot}}$$

In our analysis, we use both seed and bootstrap-based resampling to yield a total of $R = SB$ replications per language. While the B resamples may be much cheaper to obtain than the S model re-instantiations in terms of time cost, we feel that using both is helpful to properly estimate the decomposition as well as to understand whether the sources contribute equally to the total within-language variance for each language.

Why care about variance components? Most immediately, estimates of η_l (total within-language variability), σ_l (model-side variability), and τ_l (test data variability) provide the researcher with estimates of what is driving the variation in their data. High values of σ_l^2 (model-side variability) suggest that the learned distribution over responses may have high entropy or, in the case of finetuning, that the architecture may be highly sensitive to random

shuffling of batches due to e.g. small amounts of training data. On the other hand, high values of τ_l^2 (test data variability) and ν_m^2 (between-language variance) indicate that a model's performance is particularly sensitive to the exact composition of a test set or benchmark and may suggest lower ability to generalize due to overfitting. Variance components are also tied to statistical significance: we can, roughly speaking, judge an estimate as being statistically significant if it lies more than two standard errors from zero (e.g. Gelman, 2018).

Estimating model and data-side SD Using the predictions of S replications on the original test data (i.e. no bootstrapping), we estimate σ_l using the sample standard deviation formula. Since we have observations of each of the seeds on all B datasets, we also compute an estimate of the standard error $\hat{\text{se}}(\hat{\sigma}_l)$ using the sample standard deviation formula computed over the bootstrap datasets.

To estimate the boot-to-boot variability τ_l, we first compute the F1 score variance over the B bootstrap datasets separately for each of the S seeds.

We then construct the average estimator $\bar{\tau}_l = \sum_{s=1}^{S} \hat{\tau}_{ls}/S$. The standard error of $\hat{\tau}_l$ is estimated using the usual sample standard deviation formula over seeds. This gives us the plug-in estimate of the standard error of the averaged estimator $\hat{\text{se}}(\bar{\tau}_l) = \hat{\text{se}}(\hat{\tau}_{ls})/\sqrt{S}$.

Aggregate scores and their standard errors In most leaderboard-style scenarios, aggregated scalar performance measures such as the arithmetic mean $\bar{x}_{1:L}^{(m)} = \sum_l x_l^{(m)}/L$ tend to be used instead of the full score vectors. Its popularity of the arithmetic mean can be explained by how simple it is to compute, as well as the closed-form expression for its standard deviation $\text{sd}(x_1^{(m)}, \ldots, x^{(m)_L})/\sqrt{L}$ using only the between-language SD ν_m and the within-language SDs η_{ml}. Other aggregation functions, such as the geometric mean, or median may also be used, although they may not be as well-behaved in terms of standard error. For such aggregation functions, the SD typically does not exist in closed form and must be estimated using resampling.

4 Task 1: Question answering with LLMs

As our first case study, we focus on multilingual question answering and evaluate four LLMs on all subsets of XQuAD (Artetxe et al., 2020). Specifically, we use `aya-expanse-8B` (Dang et al., 2024), `TowerInstruct-Mistral-7B-v0.2` (Alves et al., 2024), Google's `gemma2-9b` and finally `Clarus-7B-v0.3`. We evaluate using token-level F1 score, using the standard `lm-evaluation-harness` implementation. Details of the corpus, experimental settings and hardware used are in Section B.1 in the Appendix.

4.1 Variance components

The `lm-evaluation-harness` library we use to run our experiments automatically computes bootstrapped standard errors for F1. In addition to the bootstrap SE, we incorporate model-side uncertainty by sampling 5 answers for each question from each model. The summarized variance components are displayed in Table 1. The full decomposition is available in Table 10 in the Appendix.

Model	Mean	SD	Min	Max
Between-language (ν)				
Clarus 7B	8.92	-	-	-
TowerInstruct 7B	11.03	-	-	-
Aya Expanse 8B	12.01	-	-	-
Gemma 2 9B	5.05	-	-	-
Total within-language (η_{ml})				
Clarus 7B	0.89	0.41	0.40	1.67
TowerInstruct 7B	0.73	0.32	0.19	1.21
Aya Expanse 8B	1.47	0.18	1.18	1.73
Gemma 2 9B	1.21	0.16	0.93	1.51
Boot-to-boot (τ_{ml})				
Clarus 7B	0.70	0.29	0.24	1.27
TowerInstruct 7B	0.59	0.28	0.15	1.05
Aya Expanse 8B	1.24	0.11	1.06	1.42
Gemma 2 9B	0.95	0.10	0.80	1.18
Seed-to-seed (σ_{ml})				
Clarus 7B	0.53	0.33	0.16	1.25
TowerInstruct 7B	0.41	0.18	0.12	0.74
Aya Expanse 8B	0.76	0.30	0.25	1.23
Gemma 2 9B	0.74	0.20	0.48	1.09

Table 1: Summary of variance components for QA experiments.

4.2 Resampling for model comparison

We seek to quantify the uncertainty in the XQuAD F1 scores in each language as well as aggregated across languages. Specifically, we compute pairwise model differences on each language as well as between aggregate scores computed using the arithmetic mean, geometric mean and median. We also compute rankings as point estimates.

We estimate the difference between models using the following parametric bootstrap resampling procedure. Given a language, we first take an esti-

	Clarus 7B			TowerInstruct 7B		Aya Exp. 8B
	TowerInstruct 7B	Aya Exp. 8B	Gemma 2 9B	Aya Exp. 8B	Gemma 2 9B	Gemma 2 9B
Arabic	8.75 ± 1.13	-24.27 ± 1.34	-0.11 ± 1.50*	-33.02 ± 1.39	-8.86 ± 1.53	24.16 ± 1.71
Chinese	-3.96 ± 1.50	-22.80 ± 1.66	-6.34 ± 1.47	-18.83 ± 1.84	-2.38 ± 1.67*	16.45 ± 1.82
English	5.21 ± 1.67	-33.30 ± 1.86	8.79 ± 1.95	-38.50 ± 1.67	3.58 ± 1.75	42.09 ± 1.95
German	-6.43 ± 1.41	-32.16 ± 1.44	-9.04 ± 1.46	-25.73 ± 1.72	-2.61 ± 1.74*	23.12 ± 1.77
Greek	16.45 ± 1.12	-20.77 ± 1.69	1.02 ± 1.58*	-37.22 ± 1.59	-15.43 ± 1.47	21.79 ± 1.94
Hindi	20.43 ± 0.94	-26.75 ± 1.51	1.26 ± 1.46*	-47.18 ± 1.34	-19.17 ± 1.27	28.01 ± 1.76
Romanian	2.36 ± 1.45*	-30.14 ± 1.55	-3.32 ± 1.81*	-32.50 ± 1.69	-5.69 ± 1.95	26.81 ± 2.01
Russian	-6.20 ± 1.59	-20.23 ± 1.65	5.36 ± 1.91	-14.02 ± 1.65	11.57 ± 1.92	25.59 ± 1.96
Spanish	8.01 ± 2.05	-17.36 ± 1.98	12.55 ± 2.23	-25.37 ± 1.70	4.54 ± 1.99	29.91 ± 1.92
Thai	21.01 ± 1.23	-3.83 ± 1.83	6.42 ± 1.69	-24.84 ± 1.52	-14.60 ± 1.35	10.25 ± 1.90
Turkish	7.06 ± 0.90	-25.13 ± 1.50	-7.01 ± 1.38	-32.19 ± 1.53	-14.07 ± 1.40	18.12 ± 1.85
Vietnamese	3.23 ± 1.06	-29.35 ± 1.41	-5.82 ± 1.44	-32.59 ± 1.39	-9.05 ± 1.41	23.53 ± 1.68

Table 2: Pairwise differences between models on XQuAD. Non-significant differences are indicated with an asterisk.

mate of each model's average performance on it by averaging the $S = 5$ observed values. We then use the estimate of the within-language SD, η_{ml}, to get a randomly resampled performance score for each model.

We estimate η_{ml} by first computing the standard deviation of scores across seeds and then use the average bootstrap SD τ_{ml} across the $S = 5$ values computed by `lm-evaluation-harness` to form an aggregate SD estimate of η_{ml}. We then sample a noise term using a random draw from $\varepsilon \sim \mathcal{N}(0, \eta_{ml}^2)$ yielding a resampled score $x + \varepsilon$ with variance $\eta_{ml}^2 = \sigma_{ml}^2 + \tau_{ml}^2$. While the score itself may not be Gaussian, we believe that modeling the noise terms as such is reasonable, as we observe mostly small score deviations across both seeds and bootstrap resamples.

We repeat this process $R = 1000$ times. For each replication, we compute a performance difference on each language, and an aggregated score of performance scores across languages, using the arithmetic mean, geometric mean and median. We also rank the models based on the randomly replicated performance scores. The results of this procedure are summarized in several tables below.

Table 2 summarizes pairwise differences between models and their associated SDs. Table 3 gives the rank distribution of each model using each of the summary statistics. While Aya Expanse 8B and TowerInstruct 7B reliably place first and last regardless of metric, the rank distributions of Clarus 7B and Gemma 2 9B can vary significantly depending on which aggregation function is used.

	Aya 8B	Clarus 7B	Gemma 2 9B	TI 7B
Arithmetic mean				
1	**100.00**	-	-	-
2	-	24.50	**75.50**	-
3	-	**75.50**	24.50	-
4	-	-	-	**100.00**
Median				
1	**100.00**	-	-	-
2	-	35.30	**64.70**	-
3	-	**64.70**	35.30	-
4	-	-	-	**100.00**
Geometric mean				
1	**100.00**	-	-	-
2	-	**94.90**	5.10	-
3	-	5.10	**94.90**	-
4	-	-	-	**100.00**

Table 3: Distribution of model ranks on XQuAD using different aggregators. Simulation computed over 1,000 randomly sampled performance scores.

5 Task 2: Machine translation with LLMs

As a second case study, we seek to compare the performance of three translation-oriented LLMs on the `devtest` split of the multilingual FLORES-200 translation benchmark (NLLB Team et al., 2024). Specifically, we use `aya-expanse-8B` (Dang et al., 2024), `TowerInstruct-Mistral-7B-v0.2` (Alves et al., 2024) (both derived from Mistral 7B), and Clarus 7B. Our reasoning for choosing these models despite the small parameter count of ∼7B is twofold. First, these models are some of the few that are either evaluated on or specifically engineered for translation. Second, our computational resources limit the size of models we can run without quantization. Ideally, we would also test on all languages covered by FLORES-200, but the language support of the LLMs restricts us to Dutch, English, French, German, Italian, Korean, Portuguese, Russian, and Spanish. For each language we create two language pairs: XX→EN and

EN→XX. We evaluate using BLEU (Papineni et al., 2002), ChrF++ (Popović, 2015) and COMET (Rei et al., 2020).

5.1 Variance components

The `lm-evaluation-harness` library we use to run our experiments automatically computes bootstrapped standard errors for BLEU, ChrF++ and COMET. In addition to the bootstrap SE, we incorporate model-side uncertainty by sampling 5 translation hypotheses for each source sentence from each model. The summarized results are shown in Table 8 in the Appendix. A detailed variance decomposition of the within-language variation is shown in Table 11 in the Appendix. As with XQuAD, the between-language variability (ν) is much larger than either the model-side or bootstrap uncertainty. Most variance components are also less than 1.0 BLEU or ChrF++ points.

5.2 Resampling for model comparison

As with XQuAD, we also wish to quantify the uncertainty in the performance differences between the two models on each language. Given observed BLEU/ChrF++/COMET scores for each language, we can easily compute a point estimate by subtraction. We quantify the uncertainty in this point estimate by resampling as with XQuAD. The estimated differences are displayed in Table 7. The most obvious observation is that Clarus 7B clearly underperforms relative to Aya Expanse 8B and TowerInstruct 7B. The latter two models perform more similarly to each other and significant differences were found in 6 out of 16 translation using BLEU. Out of the 6 significant differences, TowerInstruct 7B beat Aya Expanse 8B 4 out of 6 times. When using ChrF++, significant differences were found in 10 out of 16 translation tasks, with TowerInstruct 7B leading all of them.

6 Task 3: NER

As a third task, we show how resampling can be used for model comparison in multilingual NER when finetuning pretrained language models. Our analysis is based on the OpenNER 1.0 multilingual NER benchmark (Palen-Michel et al., 2025) which consists of 61 unique datasets covering a total of 51 languages. Since we reanalyze the results of the original paper, we use the same models: mBERT and XLM-R. For each model architecture, we vary the training data according to two conditions on each language-dataset pair: (i) `individual`, where each model is finetuned with data from only a given training set and (ii) `concatenated`, where each model is finetuned with a concatenated and downsampled version of all the training data included in OpenNER. In addition to the original results, we construct $B = 100$ bootstrap replicated data sets and reuse the same predictions from the original test set. This gives us a total of $R = SB = 1000$ replicated scores for each of the 61 datasets.

		Finetuning data	
	Model	Individual	Concatenated
ν	Glot500	8.27	6.30
	mBERT	9.95	14.38
	XLM-R	7.13	6.31
σ_l	Glot500	$1.59_{(2.51)}$	$0.61_{(0.39)}$
	mBERT	$1.15_{(2.55)}$	$0.81_{(0.45)}$
	XLM-R	$1.24_{(3.53)}$	$0.64_{(0.42)}$
τ_l	Glot500	$1.27_{(0.95)}$	$1.17_{(0.79)}$
	mBERT	$1.36_{(0.86)}$	$1.47_{(0.91)}$
	XLM-R	$1.23_{(0.84)}$	$1.18_{(0.79)}$

Table 4: Summary of variance components for NER. The three groups of rows, ν, σ and τ correspond to between-language, seed, and bootstrap SDs. Subscripts indicate standard deviations across languages.

6.1 Variance components via resampling

Like with QA and MT, we investigate how resampling-based inference can be used to better make sense of experimental variation by estimating the variance components due to model and data-side uncertainty for each of the 61 languages pairs that comprise the overall benchmark. A concise display of inferences for ν, σ and τ are shown in Table 4. A more detailed table of estimates is shown in the Appendix in Table 12.

Looking at Table 4, we see that both model and data-side variability tend to lie in the 0.6–1.6 F1 range. When we compare models in the individual versus the concatenated conditions, the models in the concatenated condition exhibit distinctly lower model-side variability. This suggests that languages with less training data benefit from larger, multilingual finetuning, even if no positive transfer learning is taking place, possibly due to a reduction in gradient instability.

Model-side variability, on the other hand, remains relatively constant across individual and concatenated conditions which suggests that the size of finetuning sets may not help much in terms of dataset-to-dataset generalization. This observation also highlights how the model and data-side vari-

ance components measure fundamentally different aspects of performance variability and how both are needed for thorough performance assessment.

6.2 Resampling for model comparison

As with QA and NMT, we also estimate quantities other than variance components. To estimate uncertainties for model rankings and pairwise differences between models, we resample the $R = SB = 1000$ replications within each language. Unlike our QA and MT experiments, we do this entirely nonparametrically by sampling a seed and replication index at random. We also show how the between-language score variability (ν) can be incorporated into the analysis by also resampling what languages are considered when estimating overall performance.

Pairwise differences To obtain estimates of the pairwise differences of average performance between models, $\Delta_{mn} = \hat{\mu}^{(m)}_{\text{arith}} - \hat{\mu}^{(n)}_{\text{arith}}$, we compute a pairwise difference matrix for each replication r. Given the $M \times M \times R$-dimensional array of pairwise differences, we average over the R-dimension to obtain a single $M \times M$ table of average differences. This is displayed as the top table in Table 5.

To account for uncertainty, we also compute a standard deviation of each pairwise difference across the R-dimension. We then divide the estimated mean difference by the SD to obtain an estimate of the effect size which allows us to assess the statistical significance of the observed differences. These effect sizes are displayed in the second part of Table 5. Except for the differences between XLM-R (ind.) and Glot500 (ind.), as well as XLM-R (conc.) and XLM-R (ind.), all estimates appear to be significantly different from zero.

We also conduct another simulation where we subsample 10 datasets without replacement and otherwise estimate effect sizes as above. The motivation for this is to incorporate between-language variability into the estimation of overall performance. Intuitively, this also corresponds to a simulating the effect of downstream usage of the tested models on smaller sets of tasks. These results are shown in the bottom part of Table 5. Overall, the effect size estimates are much smaller, with many shrinking below magnitude 2 due to the increased standard error. This suggests that using fixed benchmarks may understate the differences between two models . While the chosen threshold for significance is arbitrary, the pattern of effect size shrinkage due to the increased variability is robust.

Ranks To estimate a distribution for model rankings, we recompute the model ranks using the arithmetic mean across languages for each replication. This yields an empirical distribution estimate of model ranks which is displayed in Table 6. The bolded elements indicate the ranks observed using the original data. Overall, it seems like the ranks are stable for most models, except for Glot500 (ind.) and XLM-R (ind.) which compete for the 3rd place. We also repeated the simulation with fixed languages but observed that the ranks remained constant. This is in line with the observation from Section 6.1 that the between-language variation comprises by far the largest variance component. We suspect that under the "subsampling" condition, the ranks would display a lot more variability but leave this for future work.

7 Conclusion

In this paper, we have shown how replication and resampling can be helpful tools in the evaluation of multilingual NLP models. Through experiments on question answering, machine translation, and named entity recognition, we showed how resampling-based inference can be used to estimate different *variance components* which describe how robust model performance may be to model-related randomness as well as test set composition and between-language variation. We also showed that the between-language component may dominate the standard error of estimators such as the arithmetic mean in leaderboard-style evaluation settings that involve multiple languages/tasks.

In addition to variance components, we also showed how resampling can be used to estimate the distributions and standard errors of quantities with no closed form expressions, such as pairwise differences between models and rank distributions. We also explained how standard errors relate to statistical significance and showed that underestimated SEs may lead to overly optimistic results that ultimately fail to replicate.

We also wish to stress stress that the techniques introduced in this paper need not translate to significant additional computational overhead. In particular, between-language/task variance can be computed without any resampling and within-test set bootstrap resampling procedures can be run in seconds on a modern consumer CPU. While model-side variance may require comparatively more re-

Mean: $\bar{\Delta}$					
Model	Glot500 (i)	mBERT (c)	mBERT (i)	XLM-R (c)	XLM-R (i)
Glot500 (c)	2.01	12.24	8.63	0.79	2.00
Glot500 (i)		10.23	6.63	-1.22	-0.01
mBERT (c)			-3.60	-11.45	-10.24
mBERT (i)				-7.84	-6.63
XLM-R (c)					1.21
Effect size: $\bar{\Delta}/\text{se}(\bar{\Delta})$					
Original	Glot500 (i)	mBERT (c)	mBERT (i)	XLM-R (c)	XLM-R (i)
Glot500 (c)	3.97	41.70	18.73	2.43	3.07
Glot500 (i)		23.57	10.54	-2.38	-0.17
mBERT (c)			-9.99	-38.85	-19.21
mBERT (i)				-16.49	-9.61
XLM-R (c)					1.78
Subsampled	Glot500 (i)	mBERT (c)	mBERT (i)	XLM-R (c)	XLM-R (i)
Glot500 (c)	1.35	3.61	4.14	1.15	1.22
Glot500 (i)		2.52	3.25	-0.82	-0.00
mBERT (c)			-0.73	-3.22	-2.42
mBERT (i)				-4.05	-3.31
XLM-R (c)					0.81

Table 5: Means and effect sizes of pairwise F1 score differences $\Delta_m = \hat{\mu}^{(m)}_{\text{arith}} - \hat{\mu}^{(n)}_{\text{arith}}$.

	Glot500		mBERT		XLM-R	
	Conc.	Ind.	Conc.	Ind.	Conc.	Ind.
1	**99%**	-	-	-	1%	-
2	1%	1%	-	-	**97%**	1%
3	-	44%	-	-	2%	**54%**
4	-	**55%**	-	-	-	45%
5	-	-	1%	**99%**	-	-
6	-	-	**99%**	1%	-	-

Table 6: Bootstrap-resampled frequency distribution of ranks for each model using arithmetic mean as an aggregation function.

sources in some cases, training on multiple seeds is often only relevant when working with smaller, non-LLM models. With LLMs, it is sufficient to sample multiple responses for a given input which can be done at the fraction of the cost of an entire retraining run.

It is our hope that this paper can stimulate further research into understanding the sources, extent, and nature of performance variation in empirical NLP research. We conclude with a vision for uncertainty-aware multilingual/multitask evaluation in NLP research. Researchers should:

1. Thoughtfully consider all relevant sources of randomness when evaluating and comparing different models against each other. Draw more than one set of predictions to understand how variable the performance of a model is, and use bootstrapping to complement model-side uncertainty with data-side uncertainty estimates.

2. In multilingual/multitask evaluation setups, strive to understand how observed differences between models vary from task to task. Incorporate estimates of between-language variation into statistical comparisons to avoid false positives that may not be replicable.

3. If model performance or observed differences are very noisy, communicate this clearly and describe the differences between tasks. This is preferable to attempting to distill results to a binary judgment about which model is better. If feasible, show example cases where models agree/disagree.

4. Use resampling with the replications obtained from the above steps to estimate distributions and associated uncertainty for downstream quantities of interest, such as pairwise differences and model rankings.

Limitations

While the total number of languages we investigate is quite large, the languages are not evenly distributed across the tasks we study. Due to the length limit of this venue and the limitations of existing work, we are only able to present analysis of three tasks in this paper. We have chosen tasks with 100% human-generated annotation because we believe the results could be skewed if

we include datasets generated via system output (e.g. Pan et al., 2017), including LLM-generated annotation or automatically translated datasets.

Ethical considerations

As a position paper, this paper argues for more careful evaluation of NLP experiments, particularly when multiple languages and models are studied in parallel. We believe this can be beneficial to the larger NLP community as it may help practitioners avoid drawing poorly supported conclusions and thus avoid making non-generalizable claims about research findings. That said, our methods are slightly more involved than traditional methods (e.g. hypothesis testing) and may theoretically lead to more conservative statistical analysis which may in theory obscure real effects.

Acknowledgments

This work was supported by the grant *Improving Relevance and Recovery by Extracting Latent Query Structure* by eBay to Brandeis University. This work was also supported by Brandeis University through internal research funds.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Steven Bethard. 2022. We need to talk about random seeds. *Preprint*, arXiv:2210.13393.

Robert E. Blackwell, Jon Barry, and Anthony G. Cohn. 2025. Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores. *Preprint*, arXiv:2410.03492.

Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Vagelis Papalexakis, and Hua Wei. 2025. Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses. *Preprint*, arXiv:2502.16820.

Zizhang Chen, Pengyu Hong, and Sandeep Madireddy. 2024. Question Rephrasing for Quantifying Uncertainty in Large Language Models: Applications in Molecular Chemistry Tasks. *Preprint*, arXiv:2408.03732.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. 2025. Understanding the Uncertainty of LLM Explanations: A Perspective Based on Reasoning Topology. In *Second Conference on Language Modeling*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The Benchmark Lottery. *Preprint*, arXiv:2107.07002.

Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(1):1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10:1895–1923.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

B Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Bradley Efron. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Andrew Gelman. 2005. Analysis of variance: Why it is more important than ever. *The Annals of Statistics*, 33(1):1–31.

Andrew Gelman. 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, 44(1):16–23.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Julian P. T. Higgins, Simon G. Thompson, and David J. Spiegelhalter. 2008. A Re-Evaluation of Random-Effects Meta-Analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172(1):137–159.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zoey Liu and Bonnie Dorr. 2024. The effect of data partitioning strategy on model generalizability: A case study of morphological segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2851–2864, Mexico City, Mexico. Association for Computational Linguistics.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. 2025. Statistical Uncertainty Quantification for Aggregate Performance Metrics in Machine Learning Benchmarks. *Preprint*, arXiv:2501.04234.

Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. John Wiley & Sons.

Chester Palen-Michel, Maxwell Pickering, Maya Kruse, Jonne Sälevä, and Constantine Lignos. 2025. Open-NER 1.0: Standardized open-access named entity recognition datasets in 50+ languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33637–33662, Suzhou, China. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. 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. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Donald B. Rubin. 1981. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Shavrina Tatiana and Malykh Valentin. 2021. How not to Lie with a Benchmark: Rearranging NLP Leaderboards. *Preprint*, arXiv:2112.01342.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Rob van der Goot, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning in natural language processing research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2673–2692, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shravan Vasishth and Andrew Gelman. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, 59:1311–1342.

Nico Wagner, Michael Desmond, Rahul Nair, Zahra Ashktorab, Elizabeth M. Daly, Qian Pan, Martín Santillán Cooper, James M. Johnson, and Werner Geyer. 2024. Black-box Uncertainty Quantification Method for LLM-as-a-Judge. *Preprint*, arXiv:2410.11594.

Jiannan Xiang, Huayang Li, Yahui Liu, Lemao Liu, Guoping Huang, Defu Lian, and Shuming Shi. 2022. Investigating data variance in evaluations of automatic machine translation metrics. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 150–157, Dublin, Ireland. Association for Computational Linguistics.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via Uncertainty Quantification. In *Advances in Neural Information Processing Systems*, volume 37, pages 15356–15385. Curran Associates, Inc.

A Experimental settings

A.1 Question answering

We evaluate four LLMs on all XQuAD subsets: `aya-expanse-8B` (Dang et al., 2024), `TowerInstruct-Mistral-7B-v0.2` (Alves et al., 2024), Google's `gemma2-9b` and finally `Clarus-7B-v0.3`. All models are taken from the Open LLM Leaderboard[2] from HuggingFace.

[2] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Details of the experimental settings are in the Appendix. All experiments are run using the `lm-evaluation-harness` library (Gao et al., 2024) on A40 40GB GPUs using vLLM as an inference back-end (Kwon et al., 2023). We evaluate using token-level F1 score, using the standard `lm-evaluation-harness` implementation. We do not consider exact match as it is less robust to random deviations.

A.2 Machine translation

For each language we create two language pairs: XX→EN and EN→XX. All of our experiments use use the `devtest` split of FLORES-200 which contains approximately 1,000 sentences per language. All experiments are run using the `lm-evaluation-harness` library (Gao et al., 2024) on A40 40GB GPUs. We evaluate using BLEU (Papineni et al., 2002), ChrF++ (Popović, 2015) and COMET (Rei et al., 2020).

B Datasets used

B.1 XQuAD

XQuAD (Artetxe et al., 2020) is a multilingual extension of the SQuAD v1.1 question answering benchmark (Rajpurkar et al., 2016). XQuAD includes translations 1,190 questions, originally in English, into Arabic, German, Greek, Spanish, Hindi, Romanian, Russian, Thai, Turkish, Vietnamese, and Mandarin Chinese.

B.2 OpenNER 1.0

Our analysis is based on the OpenNER 1.0 multilingual NER benchmark (Palen-Michel et al., 2025) which consists of 61 unique datasets covering a total of 51 languages. The authors experiment with three pretrained language models (PLM): mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and Glot500-m (Imani et al., 2023). Each PLM is finetuned using two experimental conditions per language. In the "individual" condition, each PLM is finetuned separately on the train split of each of the 61 datasets. In the "concatenated" setting, the train splits of all 61 datasets are concatenated[3] together before finetuning each of the three PLMs on it. The authors run each experiment 10 times using different seeds (values 42–51).

We use model outputs from the original authors with their permission. In addition to the original results, we construct $B = 100$ bootstrap replicated data sets and reuse the same predictions from the original test set. This gives us a total of $R = SB = 1000$ replicated scores for each of the 61 datasets.

[3]Due to computational considerations, before concatenation, the highest-resource datasets are downsampled.

C Additional Tables and Figures

Additional tables follow.

C.1 Variance components

Summary variance component tables Table 8 shows a detailed summary of variance components for NMT experimets.

Detailed variance component tables Table 10 shows detailed estimates for variance components for QA. Table 11 shows detailed estimates for variance components for MT. Table 12 shows detailed estimates for variance components for NER.

C.2 Ranks

Table 6 shows the marginal distributions of ranks for the NER task.

	Aya Expanse 8B		TowerInstruct 7B
Task	TowerInstruct 7B	Clarus 7B	Clarus 7B
BLEU			
de-en	**0.40 ± 1.34**	12.31 ± 1.14	11.91 ± 1.46
en-de	**-1.13 ± 1.13**	12.48 ± 0.88	13.61 ± 1.03
en-es	**-0.35 ± 0.79**	8.64 ± 0.76	8.99 ± 0.90
en-fr	**1.01 ± 1.18**	15.32 ± 1.32	14.30 ± 1.37
en-it	**1.17 ± 0.74**	11.96 ± 0.81	10.79 ± 0.80
en-ko	1.88 ± 0.49	5.15 ± 0.45	3.27 ± 0.35
en-nl	**0.40 ± 0.90**	10.05 ± 0.72	9.65 ± 0.80
en-pt	-7.16 ± 1.03	12.85 ± 1.32	20.01 ± 1.35
en-ru	**-0.25 ± 0.94**	12.31 ± 0.76	12.56 ± 0.82
es-en	**-1.29 ± 1.06**	9.41 ± 0.93	10.70 ± 0.93
fr-en	**-1.54 ± 1.33**	14.05 ± 1.13	15.59 ± 1.37
it-en	2.47 ± 1.10	11.16 ± 0.98	8.69 ± 1.21
ko-en	-4.15 ± 0.96	6.51 ± 0.81	10.66 ± 0.95
nl-en	**-1.78 ± 1.01**	10.14 ± 0.90	11.92 ± 1.09
pt-en	-5.67 ± 1.07	18.38 ± 1.26	24.05 ± 1.24
ru-en	-2.86 ± 0.95	10.01 ± 1.06	12.86 ± 1.04
ChrF++			
de-en	-2.08 ± 0.76	6.53 ± 0.94	8.62 ± 0.94
en-de	-2.03 ± 0.72	12.35 ± 0.82	14.37 ± 0.80
en-es	**-0.97 ± 0.61**	6.48 ± 0.69	7.45 ± 0.69
en-fr	**0.04 ± 0.82**	10.40 ± 0.91	10.36 ± 0.91
en-it	**0.20 ± 0.55**	11.09 ± 0.67	10.89 ± 0.61
en-ko	**-0.20 ± 0.62**	14.27 ± 0.57	14.47 ± 0.53
en-nl	**-0.44 ± 0.67**	11.50 ± 0.74	11.94 ± 0.74
en-pt	-6.49 ± 0.71	7.78 ± 0.86	14.28 ± 0.80
en-ru	**-1.36 ± 0.78**	15.33 ± 0.80	16.69 ± 0.78
es-en	-3.99 ± 0.79	5.45 ± 0.95	9.43 ± 0.79
fr-en	-3.74 ± 0.88	7.09 ± 0.92	10.83 ± 0.93
it-en	**-0.79 ± 0.76**	7.03 ± 0.94	7.82 ± 0.95
ko-en	-4.61 ± 0.63	5.52 ± 0.68	10.13 ± 0.70
nl-en	-3.74 ± 0.74	5.37 ± 1.03	9.11 ± 1.03
pt-en	-6.04 ± 0.73	9.46 ± 1.15	15.50 ± 1.07
ru-en	-4.11 ± 0.69	5.46 ± 1.05	9.58 ± 0.99
COMET			
de-en	**0.92 ± 0.50**	6.60 ± 0.79	5.68 ± 0.86
en-de	1.19 ± 0.57	16.89 ± 0.93	15.70 ± 0.89
en-es	1.19 ± 0.59	10.46 ± 0.92	9.27 ± 0.89
en-fr	**0.35 ± 0.48**	10.86 ± 0.71	10.50 ± 0.73
en-it	**0.83 ± 0.45**	14.74 ± 0.78	13.90 ± 0.77
en-ko	5.16 ± 0.73	24.62 ± 0.94	19.45 ± 0.89
en-nl	1.26 ± 0.53	19.53 ± 0.78	18.27 ± 0.83
en-pt	-1.85 ± 0.47	10.32 ± 0.86	12.17 ± 0.85
en-ru	**-0.53 ± 0.54**	17.45 ± 0.87	17.98 ± 0.85
es-en	**0.68 ± 0.62**	7.64 ± 0.81	6.97 ± 0.84
fr-en	**0.31 ± 0.55**	7.79 ± 0.83	7.48 ± 0.84
it-en	2.57 ± 0.72	7.76 ± 0.71	5.19 ± 0.89
ko-en	**-0.78 ± 0.41**	6.51 ± 0.59	7.29 ± 0.61
nl-en	**-0.68 ± 0.57**	7.58 ± 0.89	8.26 ± 0.95
pt-en	-1.81 ± 0.40	9.40 ± 1.07	11.21 ± 1.04
ru-en	**-0.57 ± 0.40**	6.33 ± 0.73	6.90 ± 0.70

Table 7: Pairwise differences between models on NMT experiments, with uncertainty quantified using bootstrap resampling and multiple seeds. The error bar corresponds ± one standard error. Nonsignificant differences are indicated in boldface.

BLEU				
	Mean	SD	Min	Max
Between-language total (ν)				
Aya Expanse 8B	7.11	-	-	-
TowerInstruct 7B	8.49	-	-	-
Clarus 7B	4.50	-	-	-
Within-language total (η_l)				
Aya Expanse 8B	0.64	0.10	0.40	0.80
TowerInstruct 7B	0.77	0.22	0.28	1.15
Clarus 7B	0.70	0.23	0.22	1.12
Seed-to-seed (σ_l)				
Aya Expanse 8B	0.34	0.10	0.15	0.55
TowerInstruct 7B	0.47	0.16	0.13	0.78
Clarus 7B	0.48	0.20	0.14	0.92
Boot-to-boot (τ_l)				
Aya Expanse 8B	0.53	0.08	0.34	0.67
TowerInstruct 7B	0.60	0.17	0.25	0.93
Clarus 7B	0.49	0.13	0.17	0.70
ChrF++				
Between-language total (ν)				
Aya Expanse 8B	8.05	-	-	-
TowerInstruct 7B	8.99	-	-	-
Clarus 7B	9.66	-	-	-
Within-language total (η_l)				
Aya Expanse 8B	0.53	0.06	0.43	0.68
TowerInstruct 7B	0.48	0.08	0.33	0.64
Clarus 7B	0.67	0.16	0.34	0.99
Seed-to-seed (σ_l)				
Aya Expanse 8B	0.28	0.09	0.11	0.46
TowerInstruct 7B	0.25	0.08	0.08	0.42
Clarus 7B	0.43	0.18	0.11	0.82
Boot-to-boot (τ_l)				
Aya Expanse 8B	0.44	0.04	0.35	0.51
TowerInstruct 7B	0.40	0.04	0.32	0.48
Clarus 7B	0.50	0.06	0.32	0.58
COMET				
Between-language total (ν)				
Aya Expanse 8B	1.21	-	-	-
TowerInstruct 7B	2.44	-	-	-
Clarus 7B	5.75	-	-	-
Within-language total (η_l)				
Aya Expanse 8B	0.36	0.07	0.26	0.55
TowerInstruct 7B	0.38	0.10	0.23	0.63
Clarus 7B	0.74	0.12	0.53	1.01
Seed-to-seed (σ_l)				
Aya Expanse 8B	0.20	0.06	0.10	0.34
TowerInstruct 7B	0.22	0.11	0.08	0.52
Clarus 7B	0.50	0.15	0.27	0.87
Boot-to-boot (τ_l)				
Aya Expanse 8B	0.30	0.06	0.22	0.46
TowerInstruct 7B	0.30	0.06	0.21	0.44
Clarus 7B	0.53	0.06	0.41	0.66

Table 8: Summary of variance components for NMT experiments.

	Mean	SD	Min	Max
Between-language total (σ)				
Glot500 (concat)	6.30	-	-	-
Glot500 (ind)	8.27	-	-	-
mBERT (concat)	14.38	-	-	-
mBERT (ind)	9.95	-	-	-
XLM-R (concat)	6.31	-	-	-
XLM-R (ind)	7.13	-	-	-
Within-language total (η_l)				
Glot500 (concat)	1.33	0.86	0.33	4.88
Glot500 (ind)	2.19	2.56	0.39	13.98
mBERT (concat)	1.70	0.99	0.24	5.42
mBERT (ind)	1.90	2.61	0.14	20.62
XLM-R (concat)	1.36	0.87	0.15	5.08
XLM-R (ind)	1.93	3.54	0.13	28.12
Seed-to-seed (σ_l)				
Glot500 (concat)	0.61	0.39	0.19	2.09
Glot500 (ind)	1.59	2.51	0.16	13.69
mBERT (concat)	0.81	0.45	0.12	2.29
mBERT (ind)	1.15	2.55	0.06	20.35
XLM-R (concat)	0.64	0.42	0.08	2.46
XLM-R (ind)	1.24	3.53	0.04	28.12
Boot-to-boot (τ_l)				
Glot500 (concat)	1.17	0.79	0.13	4.79
Glot500 (ind)	1.27	0.95	0.12	5.31
mBERT (concat)	1.47	0.91	0.21	5.33
mBERT (ind)	1.36	0.86	0.13	5.20
XLM-R (concat)	1.18	0.79	0.13	5.00
XLM-R (ind)	1.23	0.84	0.12	5.22

Table 9: Summary of variance components for NER experiments.

	Clarus 7B			TowerInstruct 7B			Aya Expanse 8B			Gemma 2 9B		
Language	σ	τ	η	σ	τ	η	σ	τ	η	σ	τ	η
Arabic	0.20	0.52	0.56	0.46	0.43	0.63	0.25	1.15	1.18	0.48	0.80	0.93
Chinese	0.36	0.55	0.66	0.52	0.63	0.82	0.65	1.28	1.43	0.51	0.98	1.10
English	0.93	1.27	1.58	0.60	1.05	1.21	0.96	1.41	1.71	0.95	1.18	1.51
German	0.57	0.54	0.78	0.74	0.81	1.10	0.79	1.31	1.53	0.69	1.05	1.26
Greek	0.36	0.74	0.82	0.17	0.33	0.37	0.71	1.07	1.29	0.97	0.89	1.31
Hindi	0.32	0.77	0.83	0.12	0.15	0.19	0.89	1.34	1.61	0.72	0.96	1.20
Romanian	0.36	0.80	0.88	0.51	0.76	0.91	0.34	1.33	1.37	0.85	1.01	1.32
Russian	0.88	0.79	1.18	0.31	0.95	1.00	0.53	1.21	1.32	0.80	0.84	1.16
Spanish	1.25	1.11	1.67	0.49	0.76	0.91	0.88	1.26	1.54	1.09	0.88	1.40
Thai	0.59	0.70	0.91	0.19	0.30	0.35	0.72	1.06	1.28	0.65	0.92	1.13
Turkish	0.16	0.37	0.40	0.40	0.41	0.57	1.23	1.21	1.73	0.67	1.00	1.20
Vietnamese	0.34	0.24	0.42	0.42	0.52	0.67	1.18	1.19	1.67	0.51	0.89	1.03

Table 10: Detailed estimates of the standard deviations due to model-related σ and bootstrap-related τ uncertainty in question answering experiments.

	Aya Expanse 8B		TowerInstruct 7B		Clarus 7B	
	σ	τ	σ	τ	σ	τ
BLEU						
German - English	0.39	0.57	0.67	0.93	0.58	0.70
English - German	0.45	0.54	0.59	0.65	0.24	0.48
English - Spanish	0.17	0.41	0.48	0.44	0.44	0.44
English - French	0.45	0.66	0.62	0.64	0.82	0.67
English - Italian	0.24	0.47	0.18	0.49	0.46	0.41
English - Korean	0.20	0.34	0.13	0.25	0.14	0.17
English - Dutch	0.37	0.45	0.48	0.49	0.22	0.37
English - Portuguese	0.23	0.67	0.40	0.65	0.92	0.65
English - Russian	0.38	0.52	0.41	0.55	0.15	0.41
Spanish - English	0.55	0.51	0.43	0.61	0.36	0.43
French - English	0.43	0.63	0.78	0.76	0.58	0.57
Italian - English	0.26	0.53	0.65	0.66	0.59	0.51
Korean - English	0.30	0.50	0.51	0.57	0.34	0.47
Dutch - English	0.15	0.55	0.59	0.60	0.52	0.47
Portuguese - English	0.46	0.61	0.32	0.68	0.79	0.59
Russian - English	0.38	0.55	0.25	0.60	0.58	0.57
ChrF++						
German - English	0.32	0.43	0.33	0.43	0.51	0.58
English - German	0.30	0.43	0.24	0.42	0.38	0.51
English - Spanish	0.24	0.36	0.29	0.32	0.35	0.41
English - French	0.33	0.47	0.39	0.44	0.47	0.54
English - Italian	0.25	0.35	0.08	0.32	0.31	0.41
English - Korean	0.26	0.38	0.18	0.38	0.11	0.32
English - Dutch	0.28	0.37	0.30	0.37	0.39	0.42
English - Portuguese	0.21	0.50	0.21	0.41	0.39	0.53
English - Russian	0.31	0.45	0.34	0.41	0.19	0.53
Spanish - English	0.46	0.50	0.11	0.40	0.44	0.51
French - English	0.34	0.51	0.42	0.48	0.40	0.56
Italian - English	0.34	0.40	0.34	0.42	0.57	0.53
Korean - English	0.14	0.41	0.20	0.41	0.16	0.50
Dutch - English	0.11	0.51	0.29	0.44	0.72	0.52
Portuguese - English	0.33	0.49	0.11	0.40	0.82	0.56
Russian - English	0.29	0.46	0.17	0.41	0.70	0.54
COMET						
German - English	0.14	0.22	0.32	0.28	0.60	0.44
English - German	0.31	0.33	0.13	0.32	0.55	0.61
English - Spanish	0.34	0.28	0.26	0.29	0.55	0.57
English - French	0.14	0.29	0.22	0.29	0.27	0.57
English - Italian	0.10	0.31	0.12	0.28	0.45	0.54
English - Korean	0.31	0.46	0.18	0.44	0.42	0.62
English - Dutch	0.15	0.29	0.24	0.34	0.42	0.57
English - Portuguese	0.19	0.30	0.21	0.23	0.59	0.54
English - Russian	0.25	0.33	0.17	0.30	0.40	0.66
Spanish - English	0.24	0.33	0.25	0.38	0.49	0.50
French - English	0.19	0.31	0.22	0.34	0.55	0.51
Italian - English	0.23	0.25	0.52	0.36	0.39	0.49
Korean - English	0.15	0.22	0.18	0.25	0.34	0.41
Dutch - English	0.11	0.30	0.36	0.30	0.69	0.46
Portuguese - English	0.20	0.26	0.08	0.21	0.87	0.51
Russian - English	0.19	0.25	0.11	0.22	0.48	0.45

Table 11: Detailed estimates of the standard deviations due to model-related σ and bootstrap-related τ uncertainty in machine translation experiments.

	Glot500				mBERT				XLM-R			
	Concat		Individual		Concat		Individual		Concat		Individual	
	σ	τ	σ	τ	σ	τ	σ	τ	σ	τ	σ	τ
$\text{ara}_{\text{AQMAR}}$	0.44	2.63	0.71	2.21	1.81	4.37	0.62	3.04	0.83	2.97	1.54	2.50
$\text{bam}_{\text{MasakhaNER 2.0}}$	0.64	1.34	2.22	1.41	1.10	1.85	0.74	1.62	0.90	1.48	0.68	1.75
$\text{bar}_{\text{BarNER}}$	2.52	7.97	7.43	10.91	4.57	13.64	2.34	13.96	1.89	7.94	4.27	10.80
$\text{bbj}_{\text{MasakhaNER 2.0}}$	0.66	2.46	1.16	2.73	1.54	2.43	1.45	2.38	1.03	2.66	0.86	3.13
$\text{cat}_{\text{AnCora}}$	0.06	0.53	0.50	0.53	0.18	0.98	0.16	0.66	0.24	0.52	0.14	0.45
$\text{cmn}_{\text{UNER Chinese GSDSIMP}}$	0.70	2.45	0.41	2.60	1.13	9.52	0.27	2.33	0.50	2.47	0.40	2.66
$\text{cmn}_{\text{UNER Chinese GSD}}$	0.73	1.99	72.48	2.45	1.25	7.35	0.22	2.04	0.63	2.11	0.28	2.62
dan_{DaNE}	1.83	4.02	2.67	4.01	2.53	5.61	1.12	3.96	1.49	4.10	0.49	2.96
$\text{deu}_{\text{GermEval}}$	0.04	0.30	0.04	0.23	0.16	0.46	0.10	0.27	0.06	0.28	0.09	0.26
$\text{ell}_{\text{elNER}}$	0.10	0.46	0.06	0.39	0.22	0.79	0.15	0.52	0.08	0.41	0.02	0.43
$\text{eng}_{\text{Tweebank}}$	2.66	3.11	61.11	3.53	0.68	3.97	3.46	4.13	2.66	3.23	0.74	2.81
$\text{eng}_{\text{UNER English EWT}}$	0.87	1.42	0.10	1.31	0.81	1.52	0.40	1.62	0.54	1.39	0.51	1.29
eus_{EIEC}	0.66	2.24	1.00	2.09	0.57	3.75	0.75	2.88	0.78	2.01	0.83	2.19
$\text{ewe}_{\text{MasakhaNER 2.0}}$	0.19	0.35	0.24	0.45	0.33	0.49	0.19	0.63	0.10	0.37	0.23	0.50
$\text{fin}_{\text{TurkuNLP}}$	0.68	1.62	1.35	1.82	0.57	2.58	0.73	2.08	0.39	1.89	0.80	1.53
$\text{fon}_{\text{MasakhaNER 2.0}}$	0.54	1.39	0.54	1.70	0.88	1.93	1.21	1.90	0.62	1.68	0.24	1.68
$\text{glg}_{\text{SLI Galician Corpora}}$	0.58	1.04	0.34	1.18	0.69	1.40	0.43	1.32	0.64	1.00	0.19	1.13
$\text{hau}_{\text{MasakhaNER 2.0}}$	0.26	0.42	0.52	0.45	0.46	0.74	1.34	0.95	0.23	0.43	0.42	0.44
$\text{hau}_{\text{MasakhaNER}}$	0.07	1.17	0.21	1.04	0.93	1.61	0.31	1.50	0.29	1.17	0.40	0.84
$\text{heb}_{\text{NEMO SPMRL}}$	0.33	2.03	1.79	3.48	0.68	4.80	3.01	3.10	0.60	2.05	0.51	2.84
$\text{heb}_{\text{NEMO UD}}$	1.45	4.60	2.55	4.69	2.86	7.36	2.30	4.51	2.73	4.42	0.90	3.78
$\text{hin}_{\text{HiNER}}$	0.09	0.02	4.34	0.01	0.01	0.04	0.00	0.02	0.01	0.02	0.00	0.01
$\text{hrv}_{\text{hr500k}}$	0.11	0.32	0.08	0.28	0.17	0.55	0.15	0.47	0.06	0.32	0.17	0.32
$\text{ibo}_{\text{MasakhaNER 2.0}}$	0.04	0.19	0.92	0.27	0.39	0.40	0.35	0.62	0.05	0.22	0.81	0.31
$\text{ibo}_{\text{MasakhaNER}}$	0.45	0.78	2.13	0.91	0.63	1.06	0.66	1.68	0.34	0.87	0.99	0.96
$\text{kaz}_{\text{KazNERD}}$	0.07	0.38	6.27	0.21	0.36	1.14	0.18	0.25	0.15	0.36	790.74	0.18
$\text{kin}_{\text{MasakhaNER 2.0}}$	0.09	0.33	0.25	0.45	0.11	0.52	0.27	0.81	0.07	0.42	0.55	0.51
$\text{kin}_{\text{MasakhaNER}}$	0.58	1.73	0.90	2.49	1.23	2.46	4.51	2.86	0.37	1.88	1.71	2.20
$\text{lug}_{\text{MasakhaNER 2.0}}$	0.10	0.32	0.06	0.33	0.25	0.44	0.44	0.56	0.20	0.32	0.10	0.42
$\text{lug}_{\text{MasakhaNER}}$	0.46	3.47	3.27	4.14	0.51	3.97	1.58	3.32	0.48	3.38	1.82	3.92
$\text{luo}_{\text{MasakhaNER 2.0}}$	0.27	0.70	0.39	0.72	0.11	0.90	0.34	0.94	0.20	0.72	0.40	0.85
$\text{luo}_{\text{MasakhaNER}}$	0.68	4.71	187.53	7.88	1.72	5.05	4.92	8.33	1.23	4.10	4.90	5.14
$\text{mar}_{\text{L3Cube MahaNER}}$	0.19	1.13	0.38	1.19	0.51	1.82	0.33	1.35	0.08	1.28	0.25	1.18
$\text{mos}_{\text{MasakhaNER 2.0}}$	0.83	2.00	1.91	2.06	0.62	2.83	1.77	2.35	0.33	1.86	1.72	2.16
$\text{nep}_{\text{EverestNER}}$	0.07	0.29	0.09	0.32	0.19	1.07	0.15	0.53	0.09	0.31	0.14	0.34
$\text{nld}_{\text{CONLL02}}$	0.08	0.33	0.19	0.38	0.95	1.07	0.16	0.51	0.04	0.37	0.14	0.35
$\text{nno}_{\text{NorNE}}$	0.15	1.31	0.30	1.28	1.91	2.74	0.52	2.03	0.22	1.33	0.70	1.29
$\text{nob}_{\text{NorNE}}$	0.14	0.92	0.77	1.06	2.95	2.33	1.01	1.73	0.30	0.86	0.55	0.99
$\text{nya}_{\text{MasakhaNER 2.0}}$	0.16	0.22	0.07	0.21	0.25	0.36	0.47	0.42	0.05	0.26	0.12	0.29
$\text{pcm}_{\text{MasakhaNER 2.0}}$	0.15	0.45	0.30	0.49	0.45	0.79	0.65	0.67	0.11	0.52	0.09	0.54
$\text{pcm}_{\text{MasakhaNER}}$	0.14	0.96	0.75	1.33	0.50	1.59	1.03	2.30	0.44	1.25	3.29	1.77
$\text{por}_{\text{UNER Portuguese}}$	0.29	0.90	0.50	0.88	0.28	1.50	0.77	1.17	0.12	0.88	0.40	0.94
$\text{qaf}_{\text{UNER Arabizi}}$	4.37	9.44	27.94	18.19	1.55	8.86	1.49	6.32	6.06	7.12	1.01	9.57
$\text{ron}_{\text{RONEC}}$	0.09	0.27	0.13	0.25	0.17	0.44	0.09	0.29	0.04	0.23	0.04	0.22
$\text{slk}_{\text{UNER Slovak SNK}}$	0.56	1.67	2.92	2.71	0.94	3.07	3.14	2.86	0.50	1.61	1.57	2.46
$\text{slk}_{\text{WikiGoldSK}}$	0.14	0.54	0.18	0.64	0.34	1.35	0.37	0.80	0.32	0.49	0.22	0.42
$\text{slv}_{\text{ssj500k}}$	0.82	22.96	1.37	28.19	0.99	28.37	3.54	27.08	0.86	24.95	2.84	27.24
$\text{sna}_{\text{MasakhaNER 2.0}}$	0.08	0.16	0.06	0.19	0.17	0.30	0.22	0.46	0.10	0.16	0.22	0.41
$\text{spa}_{\text{AnCora}}$	0.11	0.29	0.03	0.22	0.12	0.42	0.07	0.31	0.05	0.28	0.05	0.22
$\text{spa}_{\text{CONLL02}}$	0.06	0.47	0.21	0.46	0.09	0.50	0.04	0.49	0.28	0.46	0.16	0.40
$\text{swa}_{\text{MasakhaNER 2.0}}$	0.04	0.11	0.04	0.12	0.06	0.20	0.14	0.20	0.02	0.12	0.05	0.14
$\text{swa}_{\text{MasakhaNER}}$	0.20	0.83	0.27	1.20	0.23	1.26	0.54	1.51	0.26	0.84	0.26	1.12
$\text{swe}_{\text{UNER Swedish Talkbanken}}$	2.10	5.94	102.29	9.49	5.23	9.95	414.24	10.89	1.40	5.85	3.88	6.05
$\text{tha}_{\text{ThaiNNER}}$	0.07	0.31	0.06	0.27	0.08	0.40	0.06	0.48	0.06	0.31	0.06	0.27
$\text{tsn}_{\text{MasakhaNER 2.0}}$	0.35	0.64	0.65	0.76	0.50	0.93	1.74	1.21	0.48	0.76	0.78	0.88
$\text{twi}_{\text{MasakhaNER 2.0}}$	0.40	1.84	0.38	2.30	0.93	3.04	0.57	3.29	0.19	2.18	1.39	2.52
$\text{wol}_{\text{MasakhaNER 2.0}}$	0.09	0.69	0.39	0.90	0.31	1.06	0.40	1.11	0.20	0.99	1.39	1.13
$\text{wol}_{\text{MasakhaNER}}$	0.73	6.40	17.24	7.39	1.60	7.20	1.65	8.22	1.31	6.19	2.26	7.67
$\text{xho}_{\text{MasakhaNER 2.0}}$	0.12	0.37	0.11	0.43	0.20	0.71	0.20	0.90	0.17	0.39	0.44	0.46
$\text{yor}_{\text{MasakhaNER 2.0}}$	0.20	0.63	0.83	0.63	0.65	1.05	0.35	0.78	0.51	0.66	0.32	0.68
$\text{yor}_{\text{MasakhaNER}}$	0.21	2.65	8.92	2.65	0.27	2.80	0.38	2.21	0.48	2.94	1.29	2.61

Table 12: Detailed estimates of the standard deviations due to model-related σ and bootstrap-related τ uncertainty in NER experiments.