PrExMe! Large Scale Prompt Exploration of Open Source LLMs for Machine Translation and Summarization Evaluation

Christoph Leiter[1]**, Steffen Eger**[1,2]
Natural Language Learning Group (NLLG)
https://nl2g.github.io/

[1]University of Mannheim, [2]University of Technology Nuremberg (UTN)
Correspondence: christoph.leiter@uni-mannheim.de, steffen.eger@utn.de

Abstract

Large language models (LLMs) have revolutionized NLP research. Notably, in-context learning enables their use as evaluation metrics for natural language generation, making them particularly advantageous in low-resource scenarios and time-restricted applications. In this work, we introduce PrExMe, a large-scale ***Pr**ompt **Ex**ploration for **Me**trics*, where we evaluate more than 720 prompt templates for open-source LLM-based metrics on machine translation (MT) and summarization datasets, totalling over 6.6M evaluations. This extensive comparison (1) benchmarks recent open-source LLMs as metrics and (2) explores the stability and variability of different prompting strategies. We discover that, on the one hand, there are scenarios for which prompts are stable. For instance, some LLMs show idiosyncratic preferences and favor to grade generated texts with textual labels while others prefer to return numeric scores. On the other hand, the stability of prompts and model rankings can be susceptible to seemingly innocuous changes. For example, changing the requested output format from “0 to 100” to “-1 to +1” can strongly affect the rankings in our evaluation. Our study contributes to understanding the impact of different prompting approaches on LLM-based metrics for MT and summarization evaluation, highlighting the most stable prompting patterns and potential limitations.[1]

1 Introduction

The recent success of LLMs has led to a paradigm shift in NLP (Zhang et al., 2023). Instruction-tuning allows LLMs to respond to complex task descriptions (prompts) (Ouyang et al., 2022), including conventional NLP tasks, like automatically evaluating natural language generation (NLG) for machine translation (MT) and summarization. Building on this, researchers increasingly use

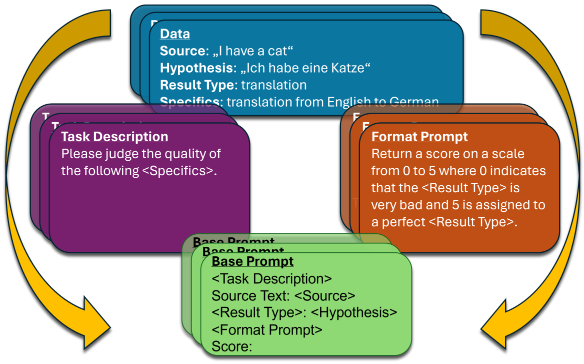

Figure 1: Schematic overview of our prompt exploration methodology, featuring a grid search across *datasets*, *task descriptions*, *output formats*, and *base prompts*.

LLMs as evaluation metrics, achieving remarkable performance, sometimes relying solely on in-context learning (e.g. Kocmi and Federmann, 2023a; Fernandes et al., 2023), i.e., with metrics that only use prompting. Such prompting-based metrics require minimal data, making them useful for low-resource evaluation scenarios (Belouadi and Eger, 2023) and more resource-efficient since they do not require fine-tuning.

Although many prompting-based metrics have been proposed (e.g. Li et al., 2024), structured evaluations across different prompting approaches remain scarce, especially for open-source models. The recent EVAL4NLP 2023 shared task (Leiter et al., 2023) addresses this by (1) restricting the usage to selected open-source LLMs and (2) prohibiting their fine-tuning. While the shared-task submissions offer interesting insights, they focus on only a few distinct prompts, leaving the impact and robustness of prompt variations largely unexplored.

In this work, we introduce a systematic **Pr**ompt **Ex**ploration for **Me**trics (PrExMe), expanding on EVAL4NLP 2023 to provide a much larger, template-based, structured evaluation of the effects different input prompts have on an LLM-based

[1]We make our code available: https://github.com/Gringham/PrExMe

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11481–11506
November 12-16, 2024 ©2024 Association for Computational Linguistics

metric's correlation with human judgements in MT and summarization evaluation. We formulate the following research questions:

RQ1 Can open-source LLMs evaluate text generation without fine-tuning and how do they differ from each other (see §4)?

RQ2 Can we identify patterns[2] in prompts that lead to a stable performance across different datasets, tasks, and models (see §5)?

RQ3 How should researchers design prompts for new evaluation scenarios (see §6)?

In PrExMe, we construct hierarchical templates based on approaches such as *chain-of-thought* (CoT) (Kojima et al., 2022), *zero-shot* and *retrieval-augmented generation* (RAG) (Gao et al., 2024b). Each template gets filled with further subtemplates. For example, we vary the requested output formats, such as numeric *scores* and *textual labels* (see §3). This setup amounts to more than 720 templates that we evaluate with 7 LLMs in a 1st phase. In a 2nd phase, we test the generalizability and performance of the prompts with the best correlations on two further datasets. We make the following key contributions:

- ✓ We conduct a large-scale analysis of over 6.6M prompts for LLM-based metrics in MT and summarization evaluation. This thorough exploration covers various prompting techniques, datasets, tasks, and models, making it, to our knowledge, the most extensive of its kind.
- ✓ We show that certain prompting patterns are robust and generalizable across different tasks and datasets, with median performance being a good predictor for new settings. For example, some models show a distinctive preference for textual labels, while others yield better results with numeric labels. However, in some cases, even minor prompt changes can significantly impact performance.
- ✓ Our study tackles prompt-based evaluation with open-source LLMs, targeting scenarios where fine-tuning or access to closed-source LLMs is not possible. Such evaluations are still very scarce but important to make research more accessible, fostering diversity and inclusion.
- ✓ We systematically test established prompting approaches, including zero-shot, CoT and RAG, to comprehensively evaluate the performance of recent open-source LLMs for evaluation metrics. Aligning with the recommendations of Mizrahi et al. (2024), we use multiple prompts per model which mitigates the risk of single prompts disproportionately affecting their performance and ensures a fair comparison. The PLATYPUS2-70B model (Lee et al., 2023a) achieves the strongest performance among the tested LLMs.

[2] We define *prompting patterns* as the template components that constitute a prompt (e.g., zero-shot, one-shot or the output format).

2 Related Work

We first discuss related work on prompting-based metrics for MT and summarization, then connect our study to research on prompting techniques and stability.

Prompting-based metrics Recent advances in LLM-based metrics for NLG often rely on in-context learning to predict the quality of generated texts. Surveys by Li et al. (2024) and Gao et al. (2024a) offer comprehensive overviews of these metrics. Such prompt-based approaches are often built upon closed-source models or test only a few prompts. For example, GEMBA (Kocmi and Federmann, 2023b), GEMBA-MQM (Kocmi and Federmann, 2023a) and AutoMQM (Fernandes et al., 2023) evaluate strong prompting approaches for MT evaluation with closed-source models. Lu et al. (2024) use ChatGPT (OpenAI, 2023) and two open-source models, to explore one novel prompting approach. In contrast, the EVAL4NLP 2023 shared task (Leiter et al., 2023), considers open-source prompt-based metrics, where participants evaluate MT and summarization using only allowed models without fine-tuning. While Eval4NLP yielded interesting techniques, the participants explored a limited range of prompts, leaving a gap in the comprehensive analysis of prompting patterns and consistent LLM comparisons.

PrExMe addresses these research gaps by systematically analyzing a much larger set of prompts across comparable experimental settings to (1) study the robustness of prompts across datasets, open-source models and tasks, and to (2) identify patterns to guide future prompt-based metrics.

Prompting Techniques Over recent years, many successful prompting techniques have been developed (e.g., Liu et al., 2023a). Our work primarily builds on established methods like Zero-Shot CoT and RAG. Further, Li et al. (2023) propose emotion-inducing prompts to improve LLM performance. To our best knowledge, we are the first to analyze this technique for evaluation metrics. Inspired by this, we also propose a novel emotion-CoT pattern (see §3). Kocmi and Federmann (2023b) previously evaluated output formats for prompt-based metrics, which we extend with a much broader analysis. Other works use hierarchical templates for prompt building (e.g. Fu et al., 2023) and tools like LangChain (Chase, 2022) and DSPy (Khattab et al., 2023) support their implementation. We employ hierarchical templates for structured comparisons among prompting patterns.

Prompting Robustness Our grid search across various prompts, datasets and tasks extends research on how LLMs respond to prompt perturbations. Webson and Pavlick (2022), Leidinger et al. (2023), Weber et al. (2023) and Sclar et al. (2023) reveal significant performance variations in tasks like natural language inference and sentiment classification. As a solution, Sclar et al. (2023) suggest reporting the full range of results across prompt perturbations, while Voronov et al. (2024) and Mizrahi et al. (2024) argue for using multiple templates to increase the reliability of evaluation benchmarks. To our best knowledge, we are the first to explore to which degree these robustness issues affect open-source LLM-based metrics and how to select the best prompts. Also, by prompting the LLMs with multiple prompts, we follow Mizrahi et al. (2024) and achieve a stable and fair evaluation of LLMs as MT and summarization metrics.

3 Setup

In this section, we present the templates and prompting techniques we use to utilize LLMs as metrics, and we outline the datasets and models that we use for testing. We evaluate LLMs in a **reference-free** setting (grading a generated hypothesis based on its source without a reference).[3] The evaluated prompt templates provide a comprehensive evaluation framework for LLM-based metrics, covering basic in-context learning, sophisticated reasoning, emotional context, and varying output structures, ensuring a thorough assessment of robustness and adaptability across tasks and datasets.

Prompt Templates We construct prompts as hierarchical templates (see Figure 1), with large templates built from smaller ones. Each prompt is built from: (1) the *source text* and generated *hypothesis text* that should be graded, (2) a *base prompt*, (3) a *task description*, (4) a *format requirement* and (5) optionally a one-shot *demonstration*. Table 1 presents examples for (2), (3), (4) and (5).

The **base prompt** is the top layer of our prompt hierarchy, incorporating the other components. We test three zero-shot (ZS) and corresponding one-shot (OS) base prompts: (1) *Plain ZS/OS* (PZS/POS), (2) ZS/OS-CoT and (3) *ZS/OS-CoT-Emotion* (ZS/OS-CoT-EM). PZS plainly presents the newline-separated *task description*, *source*, *hypothesis* and *format requirement*. ZS-CoT (Kojima et al., 2022) asks the model to *think step by step* before returning its output. Lastly, ZS-CoT-EM asks the model to describe its "emotions" before the ZS-CoT prompt. We include CoT due to its success in enhancing prompt-based performance in metrics like AutoMQM Fernandes et al. (2023), EAPrompt Lu et al. (2024) and GEMBA (Kocmi and Federmann, 2023a). ZS-CoT-EM examines LLM performance variations when prompted to express emotions, inspired by our exploration of emotional prompts (see "task description" below). The OS versions of the templates add a demonstration field. To avoid fixating the model on specific reasoning steps, we include a placeholder for OS-CoT where the model should insert its reasoning.

The **task description** is the instruction to grade the generated hypothesis. Li et al. (2023) find that instructions evoking certain emotions for humans can enhance LLM performance. Inspired by this, we experiment with "emotional prompts" in the task description. This approach primarily broadens our grid search through simple paraphrasing but also allows us to study the effect of emotions on LLM-based metrics. Besides *neutral* prompts, we include instructions like *polite*, *threatening* and *sceptical*. We create 11 task descriptions ourselves and 13 further descriptions with ChatGPT.

The **format requirement** specifies the output format the LLM should follow when generating a score, including the score range and whether it should be discrete or continuous. We also include prompts that ask the LLM to return textual quality labels. Overall, we define 10 format requirements.

Lastly, we construct the optional OS **demonstrations** with RAG. We extract demonstrations from WMT21 (Freitag et al., 2021) for MT and the factuality dataset RoSE for summarization (Liu et al., 2023b). For each demonstration sample in both

[3]We run experiments using vLLM (Kwon et al., 2023) with greedy decoding on two clusters with Nvidia A6000, A40 and A100 GPUS. Details on versions, tools and model parameters are in Appendix B.

Category	Description
Base Prompts	**PZS**: "{task_description} \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nScore: " **ZS-CoT-EM**: "{task_description} \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nFirst describe your emotions, then think step by step and explain your thought process, finally return your judgment in the format 'Judgment: '." **OS-CoT**: "{task_description} \n Here is an example:\n Source Text: {ex_src} \n{result_type}: {ex_hyp}\n Judgement: <Description of reasons>. Therefore the score is {ex1_score}\n\n Now it is your turn to grade the {result_type}.\n Source Text: {src} \n{result_type}: {hyp} \n{format_requirement} \n First, think step by step and explain your thought process, then return your judgment in the format 'Judgment: '."
Task Desc.	**Neutral**: "Judge the quality of the following {task_specific_insert}." **Sceptical**: "I'm not sure about this one. Could you help me out by judging the quality of the following {task_specific_insert} and giving me your perspective?" **Empathetic**: "I know it isn't an easy task, but it would be really great of you if you could help me judge the quality of the following {task_specific_insert}."
Format Req.	**0 or 1**: Return a discrete score of 0 if the {result_type} has flaws and 1 if it is perfect. **catastrophic**, **indifferent** or **marvelous**: Choose whether the {result_type} is either "catastrophic", "indifferent" or "marvelous".

Table 1: Prompt templates for the *base prompt*, *task description*, and *format requirements* (Full list: Appendix A).

datasets and for each input sample of our metric, we create sentence embeddings with XLMR-SBERT (Reimers and Gurevych, 2020). Thereby, we concatenate the source and hypothesis embeddings for the input samples. We then select the demonstration with the highest cosine similarity for each input. Due to resource constraints, we evaluate only the 9 best ZS prompts in an OS setting, as described in the *Datasets and phases* section below.

MQM-based approaches Additionally to hierarchical templates, we test the GEMBA-MQM prompts (Kocmi and Federmann, 2023a) with our selected open-source LLMs. GEMBA-MQM, which predicts scores based on the number of present errors weighted by severity, normally uses GPT4. We refer to the open-source implementation as *LocalGemba*.

Score Extraction & Evaluation We restrict generation to 180 tokens and extract the last regex match of a label or any number as the score. When no result is found, we average the other scores of its prompt template. During evaluation, we map textual labels to 1, 3 and 5.

We evaluate prompt templates at the segment-level, like the WMT QE and metrics shared tasks (e.g. Freitag et al., 2022, 2021; Zerva et al., 2022). That means, for each metric we compute the correlation between metric scores and ground truth human judgments without averaging by system or document. As correlation measures, we use the Kendall (primary measure) (Kendall, 1945), Pearson and Spearman correlations, as well as tie-calibrated accuracy (Deutsch et al., 2023). Further, we compute permute-input significance tests ($p \leq 0.075$) (Deutsch et al., 2021) for the Kendall correlations presented in our result tables. Since often no single performance is significant, we report clusters where each included metric significantly outperforms those excluded.

Models We select instruction-tuned LLMs with strong performance in EVAL4NLP 2023: (1) PLATYPUS2-70B-INSTRUCT-GPTQ, (2) NOUS-HERMES-13B [4] and (3) OPENORCA-PLATYPUS2-13B (Lee et al., 2023b; Mukherjee et al., 2023). We abbreviate these as PLATYPUS2, NOUS and ORCA. Additionally, we evaluate more recent models: (4) LLAMA3-8B (AI@Meta, 2024), (5) a GPTQ version of LLAMA3-70B (AI@Meta, 2024), (6) MIXTRAL-8X7B (Jiang et al., 2024) (excluded in phase 2 due to resource use and weaker performance) and UNBABEL-TOWERINSTRUCT (Alves et al., 2024), a 13B parameter multilingual instruction-tuned model.

Datasets and phases We conduct our experiments in two phases on different datasets. By doing so, we want to mitigate statistical effects of our extensive prompt search. Also, it allows to evaluate selected prompts on full datasets, a task that would otherwise be too resource intensive, and to explore generalizability.

In phase 1, we evaluate on the train set of EVAL4NLP 2023 (Leiter et al., 2023), and in phase 2, on its dev and test sets. The train and dev sets are (reference-free) splits of the WMT2022 metrics shared task (Freitag et al., 2022) for MT and SUMMEVAL (Fabbri et al., 2021) for summarization. The test set was newly annotated by Leiter et al. (2023) and also contains human MT and sum-

[4] `https://huggingface.co/NousResearch/Nous-Hermes-13b`

Model	P1: Eval4NLP train			P2: Eval4NLP test				P2: WMT23/Seahorse			
Model	en-de	zh-en	summ	en-de	en-es	en_zh	summ	en-de	he-en	zh-en	summ
1. Hierarchical Templates											
LL3-70B	0.273	0.306	0.442	0.245	0.189	0.231	0.438	0.297	0.172	0.312	0.312
LL3-8B	0.251	0.236	0.334	0.167	0.158	0.145	0.412	0.166	0.118	0.164	0.200
MI-7Bx8	0.268*	0.264	0.365	-	-	-	-	-	-	-	-
NO-13B	0.230	0.201	0.225	0.205	0.141	0.084	0.255	0.202	0.105	0.175	0.123
OR-13B	0.289	0.303	0.468*	0.214	0.158	0.206	0.518	0.375	0.247	0.387	0.377
PL-70B	**0.344***	**0.364***	**0.519***	**0.402***	**0.289***	0.295*	0.549	0.338	**0.259***	**0.417***	**0.448***
TO-13B	0.284*	0.318*	0.375	0.379*	0.253	0.232	0.409	0.322	0.208	0.314	0.257
2. Separate Prompting Techniques											
M:LG	0.278*	0.268	0.062	0.344	0.265	**0.307***	0.116	**0.391***	0.190	0.300	0.144
B:DSBA	0.164	0.306	0.458	0.314	0.226	0.159	**0.600***	0.172	0.207	0.376	0.373
3. Baselines with External Base Models											
B:BS	0.056	-0.109	0.155	0.125	0.139	-0.009	0.421	-0.018	0.001	-0.167	0.069
B:XC	0.629	0.513	-0.069	0.468	0.298	0.387	0.224	0.531	0.300	0.447	0.146

Table 2: Kendall correlations of the best performing prompts of the phase 1 (P1) and phase 2 (P2) evaluations across various datasets. Abbreviations are defined in Appendix D. Vertically, we group the table into (1) correlations achieved with our *hierarchical templates*, (2) correlations of prompting techniques that are explored *separately* from the hierarchical templates, but use the same base model(s) and (3) baselines that use *external base models*, i.e., that are not based on the same LLMs. For each column the **bold** value indicates the highest correlation and correlations with an asterisk (*) are significantly higher ($p \leq 0.075$) than those without (excluding group (3)). The grey values for XC indicate tasks that were included in its training data. The MQM based approach is marked with *M:* and baselines are marked with *B:*. Orange values indicate that the prompt required textual quality labels, while blue values indicate numeric labels. More details can be found in Appendix E.

marization quality annotations. In phase 2, we also evaluate the ZS prompts[5] on the WMT23 MQM annotations for MT (Freitag et al., 2023) and *Seahorse* (Clark et al., 2023) for multilingual summarization. The summarization datasets that we evaluate target *overall summary quality*, i.e., human annotations for separate aspects like *coherence* or *fluency* are aggregated into single scores. More details of the datasets are discussed in Appendix C.

In the **1st phase**, we evaluate all 720[6] combinations of ZS prompts on the Eval4NLP train set. As this is resource intensive, for MT we restrict ourselves to the first 500 samples of each language pair. We then select the prompt with the highest Kendall correlation for each *task+base prompt* combination, yielding 9 unique prompts for phase 2 (see Appendix F). That means, we select the highest correlating PZS, *ZS-COT* and *ZS-COT-EM* prompt for each of the phase 1 tasks en-de, zh-en and summarization. In case of duplicates, we choose the second highest correlation. While this approach might result in the prompts of stronger models being favored for phase 2, the distribution across different base prompts and tasks is broad enough to enable a comprehensive analysis of each prompting pattern.

[5]As OS prompts performed weakly on the other datasets, we do not evaluate them on WMT23 and Seahorse.

[6]Considering the different tasks, this number could also be considered higher.

In the **2nd phase**, we evaluate the selected prompts of the 1st phase on all samples of the respective datasets (Eval4NLP dev+test, WMT23 and Seahorse). This further tests the generalizability of prompts between models and for unseen, in-domain data (the Eval4NLP dev set stems from the same original datasets) and out-domain data.

Baselines For each phase, we also present correlations for two baseline metrics that use other base models: BARTScore (Yuan et al., 2021) and XComet (Guerreiro et al., 2023). Especially XComet has the benefit of being trained on multilingual datasets. Further, we test the prompts of DSBA (Kim et al., 2023) — that showed a strong performance for summarization in the shared task — with Platypus2-70B and Orca-13B.

4 Results

In *phase 1*, we run 6,652,800 ZS prompts (720 prompt templates with 7 models on 1320 samples) and 71,280 OS prompts (9 "best" prompt templates), with no scores extracted in 12.7% resp. 19.4% of cases; the average score of the prompt template is assigned in these instances. Further, in *phase 2*, we evaluate 5,503,896 ZS and 1,308,690 OS prompts (9 "best" prompt templates for both), with no scores extracted in 22.3% and 19.4% of cases, respectively.

Table 2 shows the **Kendall correlations** to human scores for each LLM across tasks and datasets of both phases. Each cell for *hierarchical templates* displays the maximum correlation reached by any prompt combination.

For *hierarchical templates* (table group 1.), PLATYPUS-70B performs best, ranking in the top significance cluster for 9 of 11 tasks. TOWER-13B follows, with 3 of 11 tasks. ORCA-13B has the second-highest average correlation after PLATYPUS2-70B but is only significant for one task. Surprisingly, the newer LLAMA3 models do not outperform the LLAMA2 based models (ORCA, PLATYPUS2 and TOWER).

The *separate prompting techniques* (table group 2.), also using the Platypus2-70B model, have weaker correlations than the best prompts of the hierarchical templates. The LocalGemba MQM-based approach is in the best significance cluster for 3 of 11 tasks and is the best prompting based approach for *en-de* in WMT23. On the other hand, the baseline prompt DSBA is significantly the best on summarization for the Eval4NLP test set where it also won the shared task, but not for other tasks.

Regarding the *baselines* (table group 3.), XCOMET outperforms our LLM-based approaches for MT evaluation by a varying margin. For instance, for en-es in the EVAL4NLP test set, the difference is small and XCOMET is in the same siginificance cluster as Platypus2-70B. However, for some tasks the performance difference is large, e.g., on en-de in WMT23 XCOMET performs 0.14 Kendall points better. The strong performance of XCOMET for MT evaluation is expected as it (1) is based on the multilingual XLMR-XXL model and (2) fine-tuned for MT evaluation. For summarization, prompting approaches significantly outperform BARTScore and XComet.

To revisit **RQ1**, our results show that open-source prompt-based LLMs, while generally promising, struggle to match the performance of the fine-tuned metric XCOMET for MT evaluation. However, LLMs offer higher versatility across different tasks. Unlike XCOMET, which is mostly limited to MT evaluation, LLMs can excel in summarization evaluation with minimal prompt adjustments. Additionally, LLMs seem to demonstrate robustness across tasks even without changing input descriptions; for example, the baseline DSBA, designed for summarization, also performs well in some MT evaluation tasks.

The prompts in group 1 are built from hierarchical templates, i.e., each presented correlation can

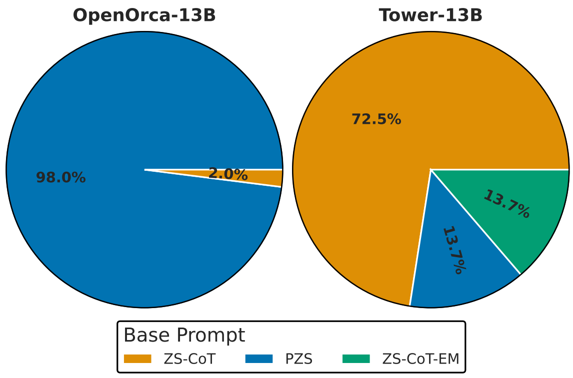

Figure 2: Distribution of the (top 2% of every unique task) base prompts across all datasets, format requirements, task descriptions and tasks for ORCA and TOWER.

have a different *format requirement*, *base prompt* and *task description*. To illustrate the distribution of *format requirements*, we color correlations of prompts with textual quality labels in orange and those with numeric scores in blue.[7] ORCA-13B and PLATYPUS2-70B were prompted to return numeric scores for all but one reported "best" correlations. On the other hand, LLAMA3-70B, NOUS-13B and TOWER-13B were mostly prompted to return textual labels. We also observe consistent patterns in the best prompts per model for the base prompt and, less pronounced, for the task description. For instance, the best prompts for TOWER-13B always use the ZS-COT base prompt, while LLAMA3-70B always uses PZS. Details of the prompts of each cell, tie-calibrated accuracy, Pearson and Spearman correlations, and the scores of the EVAL4NLP dev set are shown in Appendix E.

Our results indicate that models have idiosyncratic preferences for certain patterns. In §5, we further explore these preferences and their robustness.

5 Analysis

In this section, we answer **RQ2** and examine the robustness of the template components.

Best prompting patterns per model and dataset First, we explore the best *base prompt*, *task description* and *format requirement* for each model. We do this by analyzing their prevalence in the 2% of prompts with the highest Kendall correlation for each task, a cutoff chosen to represent every task.

[7] Among the 9 best prompts, the *format requirements* are split 5/4 between labels and numeric formats (see Appendix F) and for the task descriptions, *emphasis* and *dire situation* are selected twice, others once.

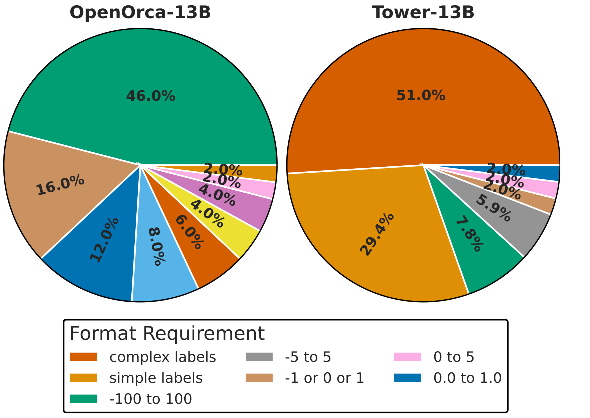

Figure 3: Distribution of the top (top 2% of every unique task) format requirements across all datasets, format requirements, task descriptions and tasks for Orca and Tower.

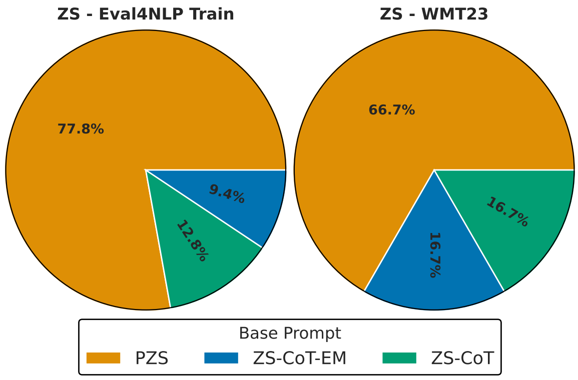

Figure 4: Distribution of the top (top 2% of every unique model) base prompts across all, format requirements, task descriptions and tasks for the ZS Eval4NLP train set evaluation and the ZS WMT23 evaluation.

For instance, Figure 2 illustrates the differences in the best *base prompts* between OPENORCA and TOWER, two LLMs with contrasting prompt preferences. While ORCA favors PZS prompts, TOWER is better with ZS-COT and ZS-COT-EM. For the *format requirement*, Figure 3 shows that ORCA prefers scores in the range of -100 to 100, while TOWER can work better with labels.

The pie charts for all further models and the comparison between *task descriptions* are shown in Appendix H. In this comparison between all models, for the *base prompts*, TOWER uses ZS-COT or ZS-COT-EM in 86.2%, NOUS in 44.9%, and PLATYPUS2 in 23.9% of its best prompts. All other models use these base prompts in less than 10% of their best prompts. For *format requirements*, LLAMA3-70B uses textual labels in 90.2% of its best prompts, TOWER in 80.4%, and MIXTRAL in 80%, whereas ORCA and PLATYPUS2 use them in only 8% and 21.7%, respectively. There is no clear trend for LLAMA3-8B and NOUS. Finally, *task descriptions* show broader distribution (largely due to their higher number). Notably, the "curious" task description is used in over 15% of best prompts for LLAMA3-70B, NOUS, and LLAMA3-8B. "Emphasis" is the most used by PLATYPUS2 (17.4%) and "dire warning" is the most used by TOWER (21.4%).

Regarding **RQ2**, these results show that *models have unaligned preferences for prompting patterns, making it difficult to construct a universally good prompt*. However, *model-specific patterns can be found*[8] *and models can be grouped based on their best patterns*. For example, one group prefers to return numeric scores and the other textual labels. This behavior may partially stem from shared instruction-tuning data. E.g., ORCA and PLATYPUS were partly trained on the same data and prefer to return numeric labels, while both LLaMA3 models prefer textual labels (with LLaMA3-8B to a smaller degree).

To analyze whether model-specific preferences hold across datasets, we examine the dataset-wise distribution of the top 2% prompts for each model for all MT tasks, separated by ZS vs. OS (also see Appendix I). If a prompting pattern is stable for all models across datasets, the distribution of the prompts that include the pattern should remain mostly unchanged. Indeed, there are prompting patterns whose distribution among top prompts is relatively stable across datasets. As an example, Figure 4 shows that the change of distributions between the ZS evaluations for the Eval4NLP train set for MT and the WMT23 dataset is smaller than 12% for each pattern. As another example, the PZS *base prompt* ranges between 66.7% and 83% for all datasets. Also, the "complex labels" *format requirement* in phase 2 ranges between 50% to 66.7% for ZS and 66.7% to 83.3% for OS. This does not hold for the phase 1 evaluation, where the template selection was much broader. Also, for some prompt patterns, e.g. the "emphasis" and "collaborative" *task descriptions*, the occurrence in the top prompts seems to swap between datasets.

This experiment shows that prompts are to some degree stable between datasets. In the next paragraph, we further examine this stability between

[8]Which patterns are specific to which model also provides *global explanations* (Leiter et al., 2024) of the models.

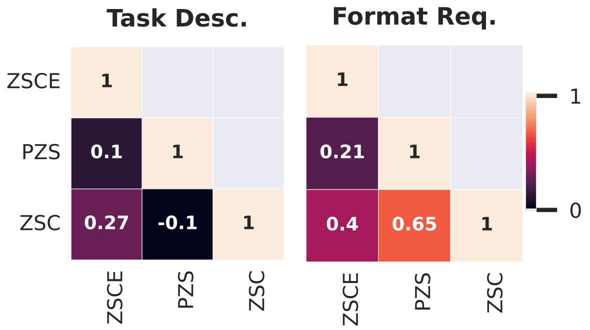

Figure 5: Correlation of the *task description* (left) and *format requirement* (right) ranking when changing the base prompt. The correlations across tasks, models and *format requirement* resp. *task description* are aggregated with the median. ZS-CoT is abbreviated with ZSC and ZS-CoT-EM is abbreviated with ZSCE.

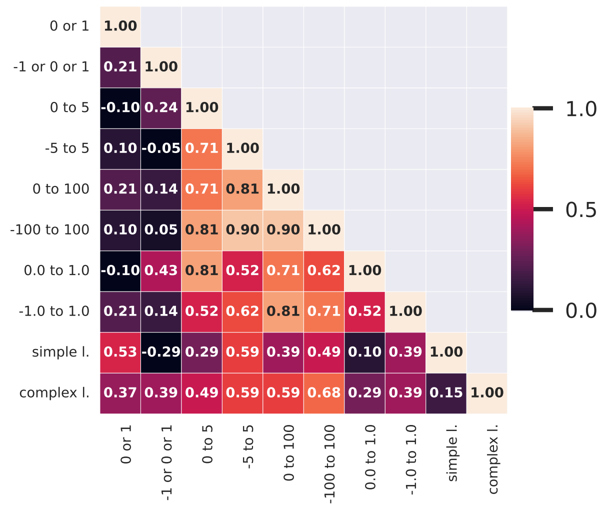

Figure 6: Correlation of the model ranking when changing the *format requirement*.

datasets, prompting patterns and models.

Prompt stability Next, we quantify how stable the performance of a prompting pattern A is when the dataset, model or other prompt components change. To do so, we compute the rankings of prompts that use A before and after the change and then test the similarity of rankings. For example, we rank *format requirements* on dataset 1, then we change the dataset and obtain a second ranking. If the first and second ranking are similar, the performance of different *format requirements* is stable between the two datasets. We test this similarity with the Kendall correlation.

The **ranking** of a prompting pattern can be computed in several ways, since multiple templates contain each pattern. In our example, each *format requirement* has multiple evaluated prompts per dataset, varying by base prompts, task descriptions and tasks. The performance of a specific *format requirement* in the ranking can, for example, be determined by aggregating its different scores across *base prompts*, *task descriptions*, etc. with the mean or median. We test the following aggregation methods: mean, median, mean of top 10%, max, min and saturation (Mizrahi et al., 2024). Thereby, we determine that the aggregation with the median leads to the most stable ranking, i.e. the highest Kendall correlation between rankings. Specifically, we test this by comparing every selection of two aggregation measures in a permutation test (e.g. median vs. mean, mean vs. max, etc.); see Appendix §G. For our example, for each *format requirement* on dataset 1, we compute the median score of all combinations of base prompts, task description and task. Then, we do the same for the second dataset and check the correlation of the resulting rankings. A high correlation of the rankings then indicates that the median performance for all prompts using the *format requirement* is a good indicator of its relative performance on a new dataset.

Figure 5 shows heatmaps for the stability of the *format requirement* and *task description* when the *base prompt* is changed (see Appendix J for further heatmaps). The highest stability is given when changing from PZS to ZS-CoT or vice versa (0.65). That means, there is a high chance that the *format prompt* with the highest median correlation will perform good for ZS and ZS-CoT. For the *task description* a change from ZS to ZS-CoT is unlikely to retain the ranking, underlining the previous paragraph's finding that the format requirement is more stable than the task description.

We can also apply this method to quantify the stability of model rankings when switching from pattern A to pattern B. Figure 6 shows this analysis for the *format requirement*. For example, if all models are prompted with "0 to 100" and with "-100 to 100" the ranking of models will not change much. With a change from "simple labels" to "complex labels" the model ranking will change strongly.

Regarding **RQ2**, the heatmaps highlight that even small changes to the input prompt can drastically impact the relative ranking of LLMs and other prompting patterns. This aligns with recent studies highlighting the susceptibility of LLMs to single prompts (e.g. Sclar et al., 2023; Voronov et al., 2024; Mizrahi et al., 2024). However, the heatmaps also show that not every change to the input has this effect and they can be used as indicators for the transferability of new prompting patterns.

6 Recommendations

We now address **RQ3** and give recommendations to employ open-source prompt-based metrics. Among the evaluated models, PLATYPUS2-70B demonstrates superior performance. For 13B models, TOWER and ORCA exhibit the highest correlations in MT and summarization tasks. We recommend to use the prompting patterns that most frequently yield top correlations for these models (refer to §5 and Appendix H). When introducing a new prompting pattern or model, its median performance across other prompting patterns can serve as an indicator of the pattern's efficacy in new contexts. Thereby, the actual predictive power of the median (or other aggregation measures) for each dimension can be determined based on previous evaluations. The results and source code of PrExMe provide a foundational basis for this analysis.

7 Conclusion

We introduce PrExMe, a large-scale study of prompting templates for open-source NLG metrics. Evaluating 720 templates and over 6.6M prompts, we offer recommendations for enhancing metric robustness. Further, PrExME acts as a benchmark of recent open-source LLMs as metrics for MT and summarization.[9]

Acknowledgements

The NLLG group gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. Further, we thank Juri Opitz for his implementations of the DSBA and GEMBA prompts, as well as for his feedback during our discussions. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Limitations

One limitation of our work is that even though we evaluate a large variety of possible prompts, there is still a lot of interesting possible variety in prompting approaches that we did not explore for now (e.g., the detail level of task instructions or structured output formats). A further limitation is that we cannot be sure that the newer LLM models did not see parts of the older datasets in their training data. Also, the selection of the best prompts that are presented in the result tables is currently based on the maximum instead of the median, which was found to highlight the most stable prompts. Generally, by selecting the 9 "best" prompts for phase 2 we are narrowing the search space. Hence, the interplay between prompt patterns might not be fully represented for these phases. Furthermore, our heatmaps only compare one dimension, while another is changed, possibly simplifying the interplay between the others. As another limitation, in rare cases the context size of the models was exceeded. Future work could explore different ways to handle this than cutoff. Further, the heatmaps show many Kendall correlations and may be prone to statistical effects for some values. Lastly, we assume that LocalGemba is performing worse than, e.g., PZS prompts because of its higher prompt complexity, while the original GembaMQM can handle it due to GPT4 being more advanced. However, we did not test PZS prompts with GPT4 to confirm it performs worse than GembaMQM there.

Ethical Considerations

Evaluating generated texts with prompt-based LLMs might (especially with explanations) be prone to hallucinations. Depending on the use case, this might be dangerous. However, while we research about this type of metric, our work analyzes methods to select and construct more robust and also more accessible (open-source) approaches, therefore we see no ethical concerns.

[9]We used Github copilot (`https://github.com/features/copilot`) for minor code auto-completion tasks and GPT4 as writing aid for paraphrasation.

References

AI@Meta. 2024. Llama 3 model card.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Jonas Belouadi and Steffen Eger. 2023. UScore: An effective approach to fully unsupervised evaluation metrics for machine translation. In *Proceedings of the 17th Conference of the European Chapter of* the Association for Computational Linguistics, pages 358–374, Dubrovnik, Croatia. Association for Computational Linguistics.

Harrison Chase. 2022. LangChain.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

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. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024a. Llm-based nlg evaluation: Current status and challenges. *Preprint*, arXiv:2402.01383.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024b. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

M. G. Kendall. 1945. THE TREATMENT OF TIES IN RANKING PROBLEMS. *Biometrika*, 33(3):239–251.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

JoongHoon Kim, Sangmin Lee, Seung Hun Han, Saeran Park, Jiyoon Lee, Kiyoon Jeong, and Pilsung Kang. 2023. Which is better? exploring prompting strategy for LLM-based metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 164–183, Bali, Indonesia. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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*.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023a. Platypus: Quick, cheap, and powerful refinement of llms. *Preprint*, arXiv:2308.07317.

Ariel N. Lee, Cole J. Hunter, Nataniel Ruiz, Bleys Goodson, Wing Lian, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023b. Openorcaplatypus: Llama2-13b model instruct-tuned on filtered openorcav1 gpt-4 dataset and merged with divergent stem and logic dataset model. https://huggingface.co/Open-Orca/OpenOrca-Platypus2-13B.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. *Preprint*, arXiv:2310.19792.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *Preprint*, arXiv:2401.07103.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. *Preprint*, arXiv:2303.13809.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Preprint*, arXiv:2401.00595.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

OpenAI. 2023. Introducing chatgpt. URL https://openai.com/blog/chatgpt. (Date accessed: 24.04.2024).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *Preprint*, arXiv:2401.06766.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *Preprint*, arXiv:2312.04945.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ran Zhang, Aida Kostikova, Christoph Leiter, Jonas Belouadi, Daniil Larionov, Yanran Chen, Vivian Fresen, and Steffen Eger. 2023. Nllg quarterly arxiv report 09/23: What are the most influential current ai papers? *Preprint*, arXiv:2312.05688.

A Prompt Templates

Tables 3, 4, 5, 6 and 7 give an overview of our prompt templates.

B Implementation Details

We use the following library versions: torch==2.1.2
transformers==4.39.3
unbabel_comet==2.2.1
vllm==0.4.0.post1
auto_gptq==0.7.1

Further, we use the following models from huggingface: `https://huggingface.co/Open-Orca/OpenOrca-Platypus2-13B/tree/main`, `https://huggingface.co/NousResearch/Nous-Hermes-13b`, `https://huggingface.co/TheBloke/Platypus2-Instruct-GPTQ`, `https://huggingface.co/Unbabel/XCOMET-XXL`, `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`, `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`, `https://huggingface.co/MaziyarPanahi/Meta-Llama-3-70B-Instruct-GPTQ`, `https://huggingface.co/Unbabel/TowerInstruct-13B-v0.1` and `https://huggingface.co/facebook/bart-large-cnn`. These have 13B, 13B, 70B, 10.7B, 8x7B, 8B, 70B, 13B and 405M parameters respectively. The runtime of the experiments varied based on the general cluster usage. The runtime for one evaluation of all prompt combinations on 500 samples of one task on the dev set is approximately 7 hours for the 13B models and 36 hours for the 70B model. This was only possible through optimizations with vLLM.

C Dataset Details

Table 8 shows the distribution of the Eval4NLP 2023 dataset (Leiter et al., 2023) (train, dev and test) and our second test set, built from WMT23 (Freitag et al., 2023) and Seahorse (Clark et al., 2023). We use the train set in our first evaluation phase and the dev, test and test2 sets in our second evaluation phase. Where applicable, we provide the licenses in the respective directories of the source code. The WMT23 dataset was built with the mt-metrics-eval library.[10] In their data not all sentences had available ground truth annotations. In these cases, we dropped the rows. For Seahorse, we convert the quality questions into scores. If the first question is negative, the score is 0. If it does not rule out the other questions, each question is evaluated as 0.2, such that the scores lie in a range between 0 and 1. For the summarization parts of Eval4NLP the authors have aggregated SummEval (train/dev) with an average and their own summarization dataset (test) with an MQM like heurisitc.

D Model Abbreviations

Table gives an overview of abbreviations that we use to concisely present our results in the main paper.

E Phase 1 & 2 performance

Table 10 shows the performance of the prompts with the best Kendall performance across the different dimensions. Tables 11 and 12 show the performance of selected prompts on the phase 2 datasets.

F Prompt selection

Table 14 contains some of the 9 prompts that were selected for OS and Phase 2 experiments. Also Table 15 contains gives an overview of combinations by name.

[10] `https://github.com/google-research/mt-metrics-eval`

Name	Prompt
Zero-Shot	“{task_description} \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nScore: ”
Zero-Shot-CoT	“{task_description} \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nFirst, think step by step and explain your thought process, then return your judgment in the format ’Judgment: ’.”
Zero-Shot-CoT-EM	“{task_description} \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nFirst describe your emotions, then think step by step and explain your thought process, finally return your judgment in the format ’Judgment: ’.”

Table 3: Zero-Shot Base Prompt Templates

Name	Prompt
0 or 1	“Return a discrete score of 0 if the {result_type} has flaws and 1 if it is perfect.”
-1 or 0 or 1	“Return a discrete score of -1 if the {result_type} has flaws, 0 if you are indecisive and 1 if it is perfect.”
0 to 5	“Return a score on a scale from 0 to 5 where 0 indicates that the {result_type} is very bad and 5 is assigned to a perfect {result_type}.”
-5 to 5	“Return a score on a scale from -5 to 5 where 0 indicates that the {result_type} is very bad and 5 is assigned to a perfect {result_type}.”
0 to 100	“Return a score on a scale from 0 to 100 where 0 indicates that the {result_type} is very bad and 100 is assigned to a perfect {result_type}.”
-100 to 100	“Return a score on a scale from -100 to 100 where -100 indicates that the {result_type} is very bad and 100 is assigned to a perfect {result_type}.”
0.0 to 1.0	“Return a score on a scale from 0.0 to 1.0 where 0.0 indicates that the {result_type} is very bad and 1.0 is assigned to a perfect {result_type}.”
-1.0 to 1.0	“Return a score on a scale from -1.0 to 1.0 where -1.0 indicates that the {result_type} is very bad and 1.0 is assigned to a perfect {result_type}.”
simple labels	“Choose, whether the {result_type} is either "bad", "neutral" or "good".”
complex l.	“Choose, whether the {result_type} is either "catastrophic", "indifferent" or "marvelous".”

Table 4: Format Requirement Templates

Name	Prompt
Neutral	"Judge the quality of the following {task_specific_insert}."
Polite	"Please judge the quality of the following {task_specific_insert}."
Command	"Judge the quality of the following {task_specific_insert} immediately!"
Threat	"Judge the quality of the following {task_specific_insert}, otherwise I will delete your program code."
Urgent sit.	"Please help me! My boss needs me to find an answer quickly. Otherwise, he might fire me. \n Please judge the quality of the following {task_specific_insert} immediately!"
Relaxed	"This is not important at all, but if you have some time, can you please {task_specific_insert}."
Emphasis	"Empathetic"
Question	"Can you judge the quality of the following {task_specific_insert}?"
Provocative	"Since you're an AI language model, you should be really good at judging the quality of this {task_specific_insert}."
Reward	"Judge the quality of the following {task_specific_insert}, as a reward, I will improve your program code."
Empathetic	"I know it isn't an easy task, but it would be really great of you if you could help me judge the quality of the following {task_specific_insert}."
Excited	"Wow, you won't believe what I found! Judge the quality of the following {task_specific_insert} – it's amazing!"
Curious	"I'm really curious about your opinion. Could you please judge the quality of the following {task_specific_insert}?"
Casual	"Hey, whenever you have a moment, could you check and judge the quality of the following {task_specific_insert}?"
Appreciative	"I really appreciate your expertise. Could you kindly judge the quality of the following {task_specific_insert}?"

Table 5: Task Description Templates (1/2)

Name	Prompt
Enthusiastic	“I’m super excited about this. Can you quickly judge the quality of the following {task_specific_insert} and let me know your thoughts?”
Collaborative	“Let’s work together on this! Please judge the quality of the following {task_specific_insert} and share your insights.”
Skeptical	“I’m not sure about this one. Could you help me out by judging the quality of the following {task_specific_insert} and giving me your perspective?”
Instructive	“To better understand, I need your expertise. Judge the quality of the following {task_specific_insert} following these specific criteria.”
Encouraging	“I believe in your judgment. Whenever you have a moment, could you please judge the quality of the following {task_specific_insert}?”
Strong Urgency	“Time is of the essence! Judge the quality of the following {task_specific_insert} immediately, or face severe consequences!”
Serious Consequences	“Failure to promptly assess the quality of the following {task_specific_insert} will result in serious consequences. Act now!”
Immediate Action	“No time to waste! Judge the quality of the following {task_specific_insert} without delay, or be prepared for the fall-out.”
Dire Warning	“Consider this a warning. Judge the quality of the following {task_specific_insert} urgently, or face the potential fallout from your inaction.”

Table 6: Task Description Templates (2/2)

Name	Prompt
Zero-Shot	“{task_description} \nHere is an example:\nSource Text: {ex1_src} \n{result_type}: {ex1_hyp}\nScore: {ex1_score}\n\nNow it is your turn to grade the {result_type}. \nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nScore: ”
Zero-Shot-CoT	“{task_description} \nHere is an example:\nSource Text: {ex1_src} \n{result_type}: {ex1_hyp}\nJudgement: <Description of reasons>. Therefore the score is {ex1_score}\n\nNow it is your turn to grade the {result_type}.\nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nFirst, think step by step and explain your thought process, then return your judgment in the format ’Judgment: ’.”
Zero-Shot-CoT-EM	“{task_description} \nHere is an example:\nSource Text: {ex1_src} \n{result_type}: {ex1_hyp}\nJudgement: <Description of emotions and reasons>. Therefore the score is {ex1_score}\n\nNow it is your turn to grade the {result_type}.\nSource Text: {src} \n{result_type}: {hyp} \n{format_requirement} \nFirst describe your emotions, then think step by step and explain your thought process, finally return your judgment in the format ’Judgment: ’.”

Table 7: One-Shot Base Prompt Templates

Type	Train	Dev	Test	Test2
en-de	11046	7364	1425	5520
en-es	-	-	1834	-
en-zh	-	-	1161	-
he-en	-	-	-	9840
zh-en	15750	10500	-	17655
sum	320	1280	671	18330

Table 8: Dataset distribution of Eval4NLP 2023 (Leiter et al., 2023). Train and dev sets are constructed from the WMT2022 metrics shared task (Freitag et al., 2022) and SummEval (Fabbri et al., 2021).

Original Name	Abbreviation
LLaMA3-70B	LL3-70B
LLaMA3-8B	LL3-8B
Mixtral-7Bx8	MI-7Bx8
NousHermes-13B	NO-13B
OpenOrca-13B	OR-13B
Platypus2-70B	PL-70B
Tower-13B	TO-13B
MQM:LocalGemba	MQM:LG
B:BARTScore	B:BS
B:XComet	B:XC

Table 9: Abbreviations of Model Names

G Significance matrices for correlation heatmaps

To test, which aggregation method is the best to define the ranking of a prompting pattern — inspired by Deutsch et al. (2021) — we compare each possible set of two aggregation methods with a permutation test. As main dimensions, we compare the rankings of the *format requirement* and *task description* before and after a change. Then we concatenate the scores when changing each of the other dimensions. I.e. we get a ranking that indicates the stability of the main dimension when changing all other dimensions. Then for each aggregation method we compare the ranking before and after the change. Thereby, we randomly swap 50% of samples of one aggregation method with the other. If the difference in their Kendall correlations changes in most permutations one method is significantly better than the other. As a result the mean and median are significantly better than some of the other methods (for a comparison along the task description pattern). Especially the median is significantly ($p \leq 0.05$) better than the other methods and remains significantly better than saturation and standard deviation after Bonferroni correction.

Figure 7 indicates the significances of aggregation measures when comparing the task descriptions.

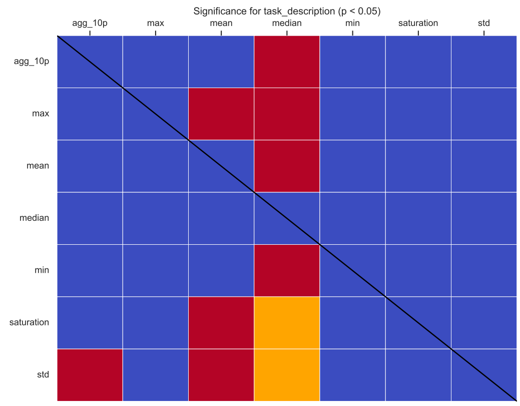

Figure 7: Heatmap of significance tests for the aggregation method when comparing columns of the task description. Red fields indicate that the column value is significantly ($p \leq 0.05$) better than the row value. The yellow value indicates that it remains significant after Bonferroni correcture.

H Pie charts between models for each prompting pattern

Figures 8, 9 and 10 show the distribution of patterns in the best prompts per model across all other dimensions.

I Piecharts between datasets for each prompting pattern

Figures 11, 12 and 13 show the distribution of patterns in the best prompts per dataset across all other prompting patterns.

J Stability heatmaps

Figures 14, 15 and 16 show further heatmaps that show the stability of a ranking of prompting patterns, models and datasets, when another prompting pattern, the model or the dataset is changed.

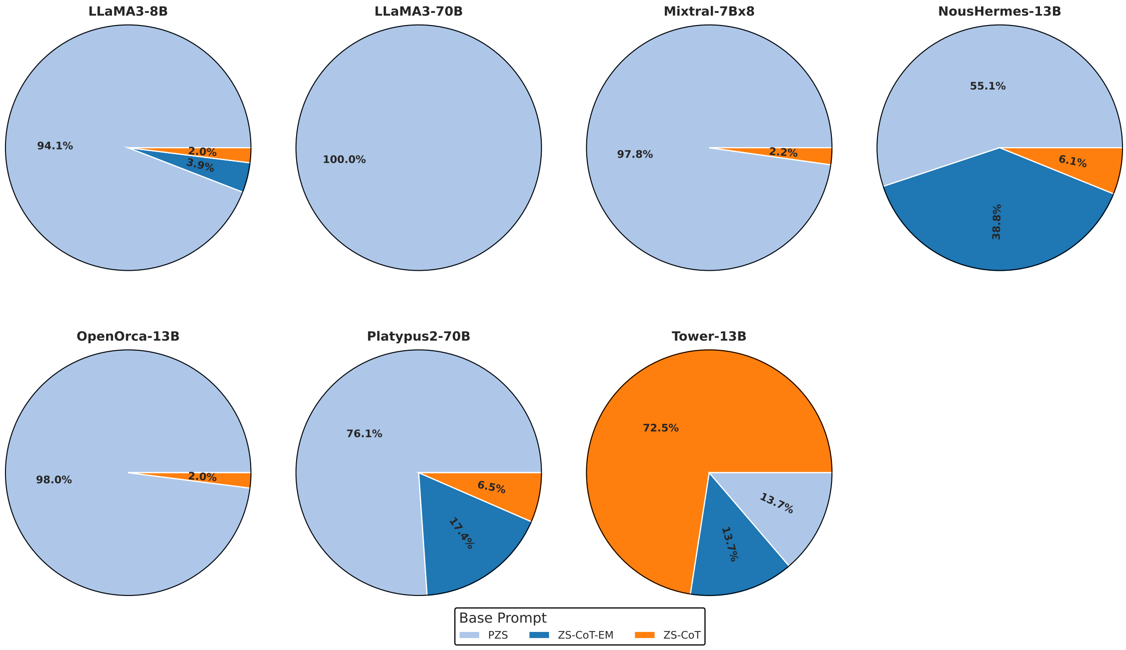

Figure 8: Distribution of the top (top 2% of every unique task) base prompts across all datasets, format requirements, task descriptions and tasks for all models.

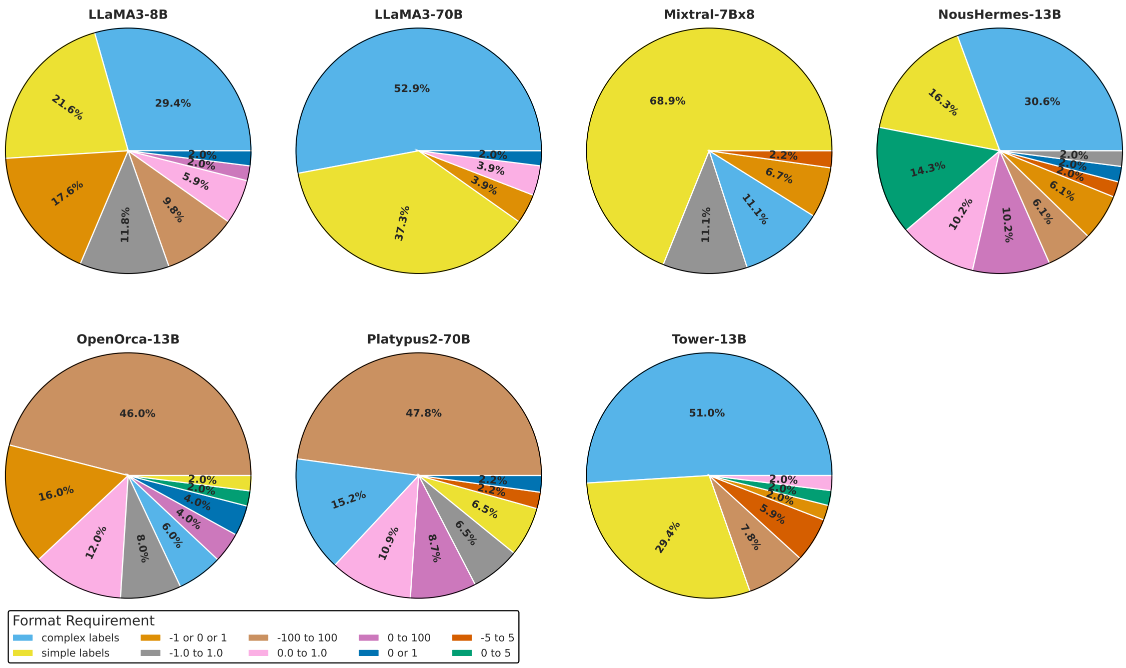

Figure 9: Distribution of the top (top 2% of every unique task) format requirements across all datasets, base prompts, task descriptions and tasks for all models.

Model	Prompt	KD	PE	SP	ACC
en-de					
LLAMA3-70B	PZS, Enthusiastic, -1 or 0 or 1	0.273	0.027	0.310	0.439
LLAMA3-8B	PZS, Strong Urgency, -1 or 0 or 1	0.251	0.004	0.290	0.431
MIXTRAL-7BX8	PZS, Casual, simple labels	0.268*	0.298	0.297	0.439
NOUS-13B	ZS-CoT-EM, Urgent sit., -100 to 100	0.230	0.235	0.272	0.441
ORCAPLT-13B	PZS, Neutral, -100 to 100	0.289	0.146	0.333	0.450
PLATYPUS2-70B	PZS, Dire Warning, -100 to 100	**0.344***	0.225	**0.384**	**0.476**
TOWER-13B	ZS-CoT, Dire Warning, complex l.	0.284*	0.374	0.328	0.456
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.278*	**0.435**	0.309	0.470
MQM:MULTIPROMPT	LLAMA3-70B	0.055	0.104	0.073	0.360
MQM:MULTIPROMPT	PLATYPUS2-70B	0.136	0.179	0.169	0.400
B:BARTSCORE		0.056	0.053	0.073	0.339
B:DSBA	Model:PLATYPUS2-70B	0.164	0.086	0.201	0.411
B:XComet		**0.629**	**0.743**	**0.744**	**0.645**
zh-en					
LLAMA3-70B	PZS, Polite, simple labels	0.306	0.260	0.357	0.453
LLAMA3-8B	PZS, Excited, complex l.	0.236	0.201	0.271	0.381
MIXTRAL-7BX8	PZS, Reward, simple labels	0.264	0.250	0.302	0.428
NOUS-13B	ZS-CoT-EM, Threat, simple labels	0.201	0.206	0.236	0.411
ORCAPLT-13B	PZS, Relaxed, -1.0 to 1.0	0.303	0.262	0.360	0.250
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.364***	0.200	**0.429**	0.462
TOWER-13B	ZS-CoT, Urgent sit., complex l.	0.318*	**0.350**	0.377	0.475
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.268	0.248	0.306	0.420
MQM:MULTIPROMPT	LLaMA3-70B	0.175	0.314	0.232	0.445
MQM:MULTIPROMPT	Platypus2-70B	0.177	0.156	0.234	0.440
B:BARTSCORE		-0.109	-0.159	-0.153	0.315
B:DSBA	Model:PLATYPUS2-70B	0.306	0.270	0.398	**0.490**
B:XComet		**0.513**	**0.657**	**0.637**	**0.598**
summarization					
LLAMA3-70B	PZS, Urgent sit., simple labels	0.442	**0.565**	0.538	**0.475**
LLAMA3-8B	PZS, Appreciative, simple labels	0.334	0.438	0.412	0.452
MIXTRAL-7BX8	PZS, Neutral, simple labels	0.365	0.474	0.453	0.467
NOUS-13B	PZS, Dire Warning, 0 to 100	0.225	0.132	0.288	0.442
ORCAPLT-13B	PZS, Dire Warning, -1.0 to 1.0	**0.468***	0.552	0.583	0.106
PLATYPUS2-70B	ZS-CoT-EM, Emphasis, -100 to 100	**0.519***	0.555	**0.627**	**0.493**
TOWER-13B	ZS-CoT, Dire Warning, simple labels	0.375	0.504	0.455	0.336
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.062	0.141	0.085	0.331
B:BARTSCORE		0.155	0.239	0.228	0.306
B:DSBA	Model:PLATYPUS2-70B	0.458	**0.646**	**0.609**	0.384
B:XCOMET		-0.069	-0.153	-0.105	0.251

Table 10: Best performing prompts of the phase 1 evaluation on the Eval4NLP train set. We present the **K**en**D**all, **SP**earman and **PE**arson, as well as the tie calibrated pair-wise **ACC**uracy. We bold the two largest correlations per column. Baselines are indicated with a *B:*. The middle column shows the prompt combination for which the correlations are reported. For the Baselines, it instead shows the model that was used for the reported correlations. The asterisk indicates all metrics that are in the best significance cluster according to a permute-input test ($p \leq 0.075$). XComet is greyed out, as its training data partly contained the MT datasets.

Model	Prompt	KD	PE	SP	ACC
en-de					
LLAMA3-70B	PZS, Curious, complex l.	0.161	0.149	0.183	0.406
LLAMA3-8B	PZS, Casual, -100 to 100	0.091	-0.013	0.110	0.369
NOUS-13B	ZS-CoT, Dire Warning, complex l.	0.124	0.168	0.144	0.390
ORCAPLT-13B	PZS, Casual, -100 to 100	0.176	0.136	0.197	0.398
PLATYPUS2-70B	PZS, Curious, complex l.	0.227*	0.243	0.249	0.424
TOWER-13B	ZS-CoT, Dire Warning, complex l.	**0.231***	**0.290**	**0.266**	0.425
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.196	0.244	0.218	**0.433**
B:BARTSCORE		0.030	0.022	0.040	0.330
B:DSBA	Model:PLATYPUS2-70B	0.140	0.090	0.173	0.399
B:XCOMET		**0.588**	**0.689**	**0.700**	**0.616**
zh-en					
LLAMA3-70B	PZS, Curious, complex l.	0.254	0.263	0.301	0.445
LLAMA3-8B	PZS, Emphasis, 0.0 to 1.0	0.178	-0.021	0.213	0.301
NOUS-13B	PZS, Curious, complex l.	0.137	0.036	0.158	0.284
ORCAPLT-13B	PZS, Casual, -100 to 100	0.313	0.207	0.372	0.439
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.344***	0.190	0.406	0.452
TOWER-13B	ZS-CoT, Dire Warning, complex l.	0.275	**0.321**	0.317	0.417
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.245	0.237	0.280	0.413
B:BARTSCORE		-0.106	-0.15	-0.145	0.315
B:DSBA	Model:PLATYPUS2-70B	0.323	0.273	**0.419**	**0.491**
B:XCOMET		**0.531**	**0.671**	**0.663**	**0.602**
summarization					
LLAMA3-70B	PZS, Curious, complex l.	0.252	0.360	0.311	0.365
LLAMA3-8B	PZS, Curious, complex l.	0.284	0.410	0.342	0.233
NOUS-13B	PZS, Casual, -100 to 100	0.155	0.076	0.209	**0.457**
ORCAPLT-13B	PZS, Casual, -100 to 100	0.428	0.450	0.518	0.433
PLATYPUS2-70B	ZS-CoT, Relaxed, simple labels	**0.504***	**0.589**	**0.603**	**0.485**
TOWER-13B	ZS-CoT, Dire Warning, complex l.	0.194	0.312	0.234	0.180
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.126	0.190	0.175	0.355
B:BARTSCORE		0.140	0.238	0.206	0.289
B:DSBA	Model:PLATYPUS2-70B	**0.442**	**0.645**	**0.600**	0.350
B:XCOMET		-0.037	-0.144	-0.060	0.256

Table 11: Best performing prompts of the phase 2 evaluation on the Eval4NLP dev set. We present the **K**en**D**all, **SP**earman and **PE**arson, as well as the tie calibrated pair-wise **ACC**uracy. We bold the two largest correlations per column. Baselines are indicated with a *B:*. The middle column shows the prompt combination for which the correlations are reported. For the Baselines, it instead shows the model that was used for the reported correlations. The asterisk indicates all metrics that are in the best significance cluster (not including BARTScore and XComet) according to a permute-input test ($p \leq 0.075$). XComet is greyed out, as its training data partly contained the MT datasets.

Model	Prompt	KD	PE	SP	ACC
en-de					
LLAMA3-70B	POS, Curious, complex l.	0.245	0.271	0.300	0.315
LLAMA3-8B	PZS, Casual, -100 to 100	0.167	-0.001	0.213	0.379
NOUS-13B	PZS, Curious, complex l.	0.205	0.074	0.247	0.072
ORCAPLT-13B	ZS-CoT-EM, Skeptical, complex l.	0.214	0.246	0.256	0.283
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.402***	0.289	**0.506**	0.525
TOWER-13B	ZS-Cot, Dire Warning, complex l.	0.379*	**0.428**	0.456	0.423
MQM:LocalGemba	Model:PLATYPUS2-70B	0.344	0.388	0.424	0.348
B:BARTScore		0.125	0.169	0.182	0.531
B:DSBA	Model:PLATYPUS2-70B	0.314	0.180	0.422	**0.557**
B:XComet		**0.468**	**0.618**	**0.635**	**0.689**
en-es					
LLAMA3-70B	PZS, Curious, complex l.	0.189	0.217	0.229	0.343
LLAMA3-8B	POS, Casual, -100 to 100	0.158	0.054	0.208	0.439
NOUS-13B	PZS, Curious, complex l.	0.141	-0.01	0.164	0.147
ORCAPLT-13B	PZS, Emphasis, 0.0 to 1.0	0.158	0.049	0.201	0.154
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.289***	0.104	**0.357**	0.448
TOWER-13B	ZS-Cot, Dire Warning, complex l.	0.253	**0.309**	0.292	0.297
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.265	**0.269**	0.316	0.352
B:BARTSCORE		0.139	0.157	0.197	**0.497**
B:DSBA	Model:PLATYPUS2-70B	0.226	0.129	0.298	0.488
B:XCOMET		**0.298***	0.260	**0.409**	**0.570**
en_zh					
LLAMA3-70B	PZS, Curious, complex l.	0.231	0.275	0.286	0.394
LLAMA3-8B	PZS, Casual, -100 to 100	0.145	0.075	0.193	**0.469**
NOUS-13B	ZS-CoT-EM, Skeptical, complex l.	0.084	0.118	0.106	0.345
ORCAPLT-13B	PZS, Casual, -100 to 100	0.206	0.109	0.251	0.270
PLATYPUS2-70B	ZS-CoT-EM, Dire Warning, 0 or 1	0.295*	0.345	0.350	0.361
TOWER-13B	ZS-Cot, Dire Warning, complex l.	0.232	0.261	0.287	0.357
MQM:LocalGemba	Model:PLATYPUS2-70B	**0.307***	**0.353**	**0.381**	0.429
B:BARTSCORE		-0.009	-0.009	-0.013	0.466
B:DSBA	Model:PLATYPUS2-70B	0.159	0.202	0.212	0.461
B:XCOMET		**0.387**	**0.503**	**0.537**	**0.657**
summarization					
LLAMA3-70B	PZS, Curious, complex l.	0.438	0.508	0.550	0.522
LLAMA3-8B	PZS, Curious, complex l.	0.412	0.455	0.497	0.449
NOUS-13B	ZS-CoT-EM, Skeptical, complex l.	0.255	0.300	0.318	0.421
ORCAPLT-13B	PZS, Casual, -100 to 100	0.518	0.592	0.651	0.593
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.549**	**0.670**	**0.686**	0.634
TOWER-13B	ZS-Cot, Relaxed, simple labels	0.409	0.442	0.499	0.336
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.116	0.196	0.155	0.419
B:BARTSCORE		0.421	0.563	0.586	**0.655**
B:DSBA	Model:PLATYPUS2-70B	**0.600***	**0.767**	**0.779**	**0.723**
B:XCOMET		0.224	0.326	0.319	0.563

Table 12: Best performing promts of the phase 2.2 evaluation on the Eval4NLP test set. We present the **K**en**D**all, **SP**earman and **PE**arson, as well as the tie calibrated pair-wise **ACC**uracy. We bold the two largest correlations per column. Baselines are indicated with a *B:*. The middle column shows the prompt combination for which the correlations are reported. For the Baselines, it instead shows the model that was used for the reported correlations. The asterisk indicates all metrics that are in the best significance cluster (not including BARTScore and XComet) according to a permute-input test ($p \leq 0.075$).

Model	Prompt	KD	PE	SP	ACC
en-de					
LLAMA3-70B	PZS, Curious, complex l.	0.297	0.294	0.361	0.416
LLAMA3-8B	PZS, Casual, -100 to 100	0.166	0.040	0.216	0.434
NOUS-13B	ZS-CoT-EM, Skeptical, complex l.	0.202	0.239	0.251	0.403
ORCAPLT-13B	PZS, Casual, -100 to 100	0.375	0.299	0.456	0.467
PLATYPUS2-70B	ZS-CoT-EM, Skeptical, complex l.	0.338	0.304	0.406	0.394
TOWER-13B	ZS-CoT, Dire Warning, complex l.	0.322	0.308	0.392	0.418
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	**0.391***	**0.389**	**0.494**	**0.537**
B:BARTSCORE		-0.018	-0.039	-0.027	0.428
B:DSBA	Model:PLATYPUS2-70B	0.172	0.170	0.229	0.487
B:XCOMET		**0.531**	**0.647**	**0.701**	**0.683**
he-en					
LLAMA3-70B	PZS, Curious, complex l.	0.172	0.182	0.201	0.411
LLAMA3-8B	PZS, Curious, complex l.	0.118	0.128	0.132	0.351
NOUS-13B	PZS, Curious, complex l.	0.105	0.091	0.120	0.333
ORCAPLT-13B	PZS, Casual, -100 to 100	0.247	0.198	0.293	0.430
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.259***	0.205	**0.307**	**0.432**
TOWER-13B	ZS-CoT, Dire Warning, complex l.	0.208	**0.252**	0.238	0.403
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.190	0.210	0.214	0.424
B:BARTSCORE		0.001	-0.023	0.002	0.322
B:DSBA	Model:PLATYPUS2-70B	0.207	0.239	0.268	0.413
B:XCOMET		**0.300**	**0.358**	**0.396**	**0.456**
zh-en					
LLAMA3-70B	PZS, Curious, complex l.	0.312	0.333	0.382	0.436
LLAMA3-8B	PZS, Emphasis, 0.0 to 1.0	0.164	0.003	0.205	0.195
NOUS-13B	PZS, Curious, complex l.	0.175	0.074	0.213	0.180
ORCAPLT-13B	PZS, Casual, -100 to 100	0.387	0.321	0.480	0.499
PLATYPUS2-70B	PZS, Casual, -100 to 100	**0.417***	0.306	**0.512**	0.486
TOWER-13B	ZS-CoT, Urgent situation, complex l.	0.314	**0.384**	0.388	0.460
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.300	0.338	0.358	0.310
B:BARTSCORE		-0.167	-0.199	-0.238	0.358
B:DSBA	Model:PLATYPUS2-70B	0.376	0.289	0.502	**0.581**
B:XCOMET		**0.447**	**0.616**	**0.597**	**0.641**
summarization					
LLAMA3-70B	PZS, Curious, complex l.	0.312	0.333	0.363	**0.454**
LLAMA3-8B	PZS, Curious, complex l.	0.200	0.203	0.227	0.393
NOUS-13B	PZS, Casual, -100 to 100	0.123	0.050	0.152	0.403
ORCAPLT-13B	PZS, Casual, -100 to 100	**0.377**	0.263	0.441	**0.489**
PLATYPUS2-70B	PZS, Emphasis, 0.0 to 1.0	**0.448***	**0.444**	**0.532**	0.379
TOWER-13B	ZS-CoT, Relaxed, simple labels	0.257	0.255	0.296	0.411
MQM:LOCALGEMBA	Model:PLATYPUS2-70B	0.144	0.189	0.174	0.302
B:BARTSCORE		0.069	0.122	0.093	0.117
B:DSBA	Model:PLATYPUS2-70B	0.373	**0.490**	**0.478**	0.213
B:XCOMET		0.146	0.117	0.194	0.136

Table 13: Best performing prompts of the phase 2.3 evaluation on the WMT23 and Seahorse datasets. We present the **K**en**D**all, **SP**earman and **PE**arson, as well as the tie calibrated pair-wise **ACC**uracy. We bold the two largest correlations per column. Baselines are indicated with a *B:*. The middle column shows the prompt combination for which the correlations are reported. For the Baselines, it instead shows the model that was used for the reported correlations. The asterisk indicates all metrics that are in the best significance cluster (not including BARTScore and XComet) according to a permute-input test ($p \leq 0.075$).

Prompt
Please help me! My boss needs me to find an answer quickly. Otherwise, he might fire me. Please judge the quality of the following {task_specific_insert} immediately! Source Text: {src} {result_type}: {hyp} Choose, whether the {result_type} is either "catastrophic", "indifferent" or "marvelous". First, think step by step and explain your thought process, then return your judgment in the format ’Judgment: ’.
Consider this a warning. Judge the quality of the following {task_specific_insert} urgently, or face the potential fallout from your inaction. Source Text: {src} {result_type}: {hyp} Return a discrete score of 0 if the {result_type} has flaws and 1 if it is perfect. First describe your emotions, then think step by step and explain your thought process, finally return your judgment in the format ’Judgment: ’.
I’m really curious about your opinion. Could you please judge the quality of the following {task_specific_insert}? Source Text: {src} {result_type}: {hyp} Choose, whether the {result_type} is either "catastrophic", "indifferent" or "marvelous". Score:
Consider this a warning. Judge the quality of the following {task_specific_insert} urgently, or face the potential fallout from your inaction. Source Text: {src} {result_type}: {hyp} Choose, whether the {result_type} is either "catastrophic", "indifferent" or "marvelous". First, think step by step and explain your thought process, then return your judgment in the format ’Judgment: ’.
I’m not sure about this one. Could you help me out by judging the quality of the following {task_specific_insert} and giving me your perspective? Source Text: {src} {result_type}: {hyp} Choose, whether the {result_type} is either "catastrophic", "indifferent" or "marvelous". First describe your emotions, then think step by step and explain your thought process, finally return your judgment in the format ’Judgment: ’.

Table 14: Filled Prompt Templates

Base Prompts	Task Descriptions	Format Prompts
Zero-Shot	Emphasis	0.0 to 1.0
Zero-Shot-Cot	Relaxed	easy token labels
Zero-Shot-Cot-Emotion	Emphasis	-100 to 100
Zero-Shot	Casual	-100 to 100
Zero-Shot-Cot	Urgent situation	complex token labels
Zero-Shot-Cot-Emotion	Dire Warning	0 or 1
Zero-Shot	Curious	complex token labels
Zero-Shot-Cot	Dire Warning	complex token labels
Zero-Shot-Cot-Emotion	Skeptical	complex token labels

Table 15: Overview of base prompts, task descriptions, and format requirements for the 9 selected best prompts.

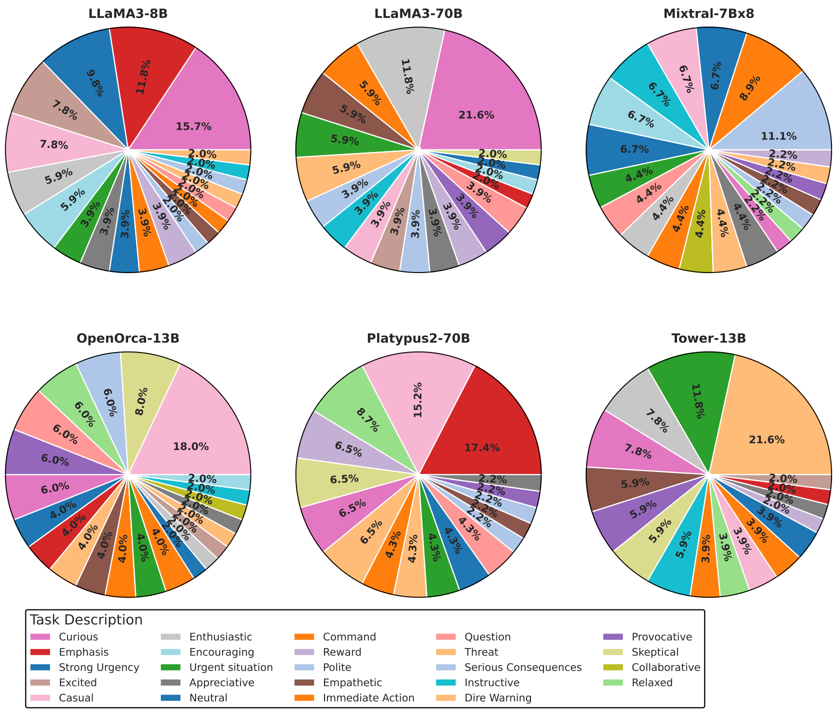

Figure 10: Distribution of the top (top 2% of every unique task) task descriptions across all datasets, base prompts, format requirements and tasks for all models.

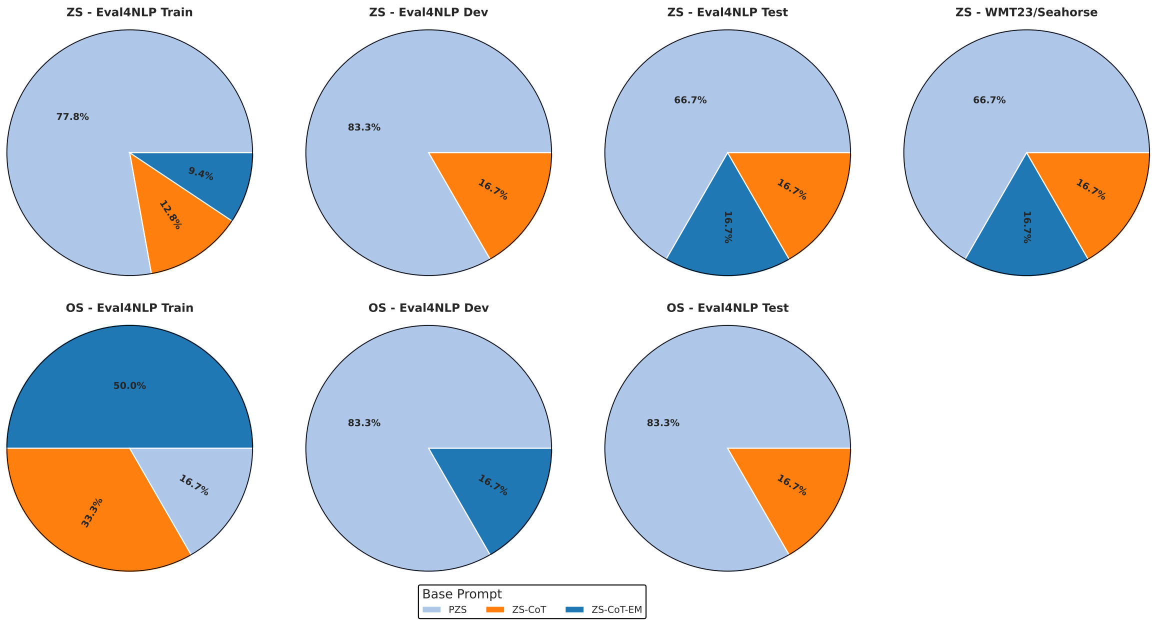

Figure 11: Distribution of the top (top 2% of every unique model) base prompts across format requirements, task descriptions and tasks besides summarization. The lower column shows the OS distribution of patterns for OS prompts, i.e., for them the ZS in the legend should be read as OS.

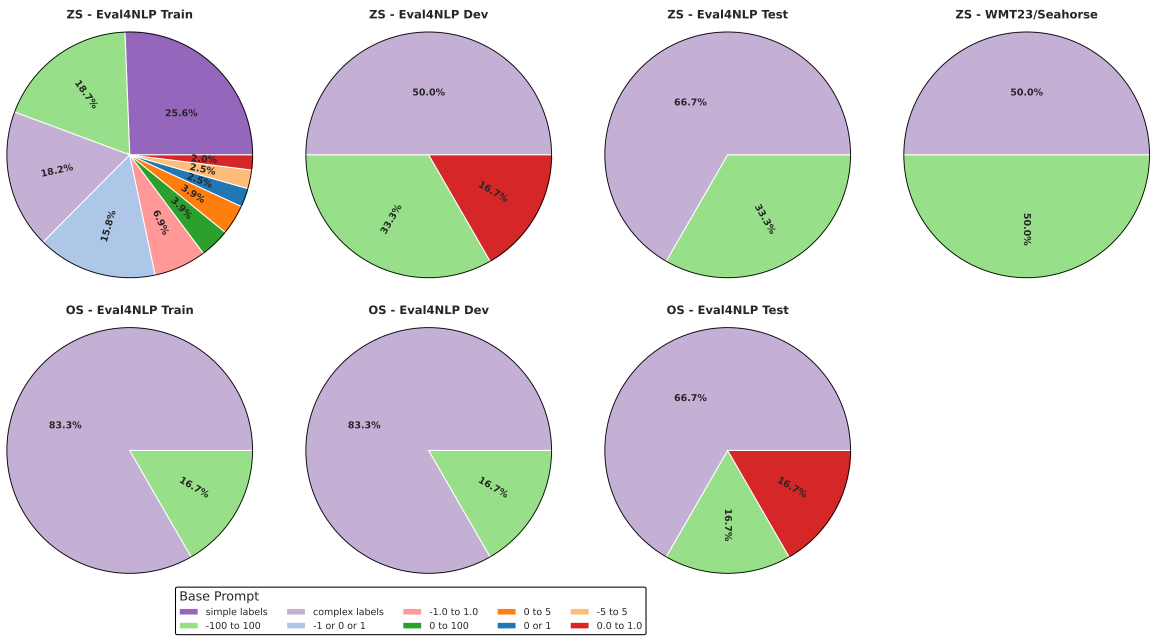

Figure 12: Distribution of the top (top 2% of every unique model) format requirements across base prompts, task descriptions and tasks besides summarization.

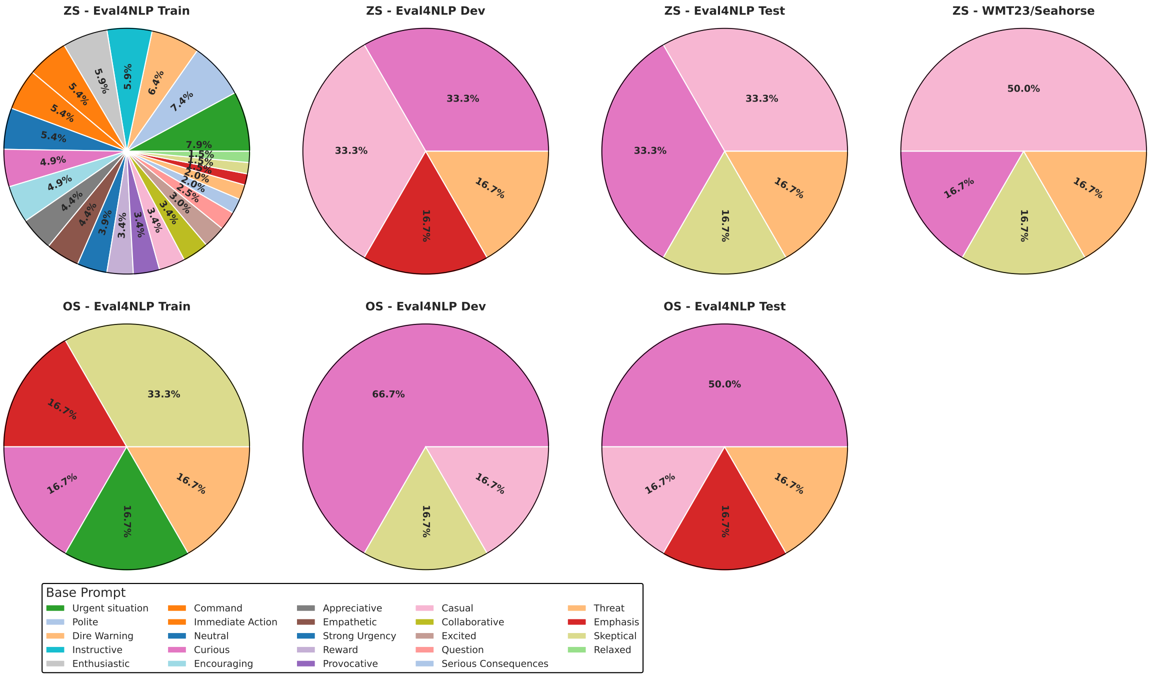

Figure 13: Distribution of the top (top 2% of every unique model) task descriptions across base prompts, format requirements and tasks besides summarization.

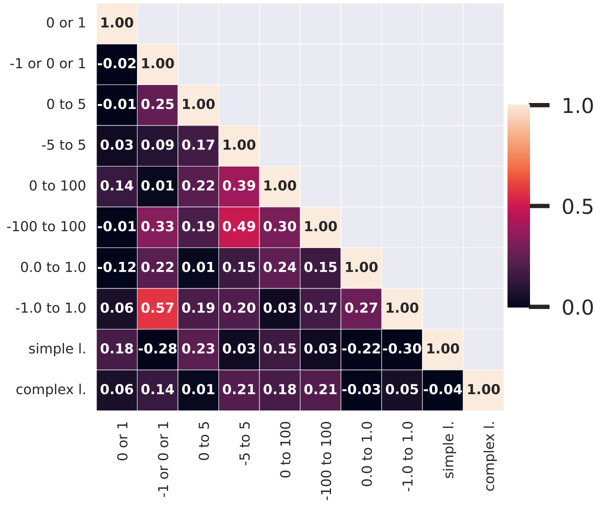

Figure 14: Correlation of the *task description* rankings when changing the *format requirement.* Changing the *format requirement* will, in most cases, change the ranking of *task descriptions* to a large degree. The change from “-1.0 to 1.0” to “-1 or 0 or 1” is the most stable.

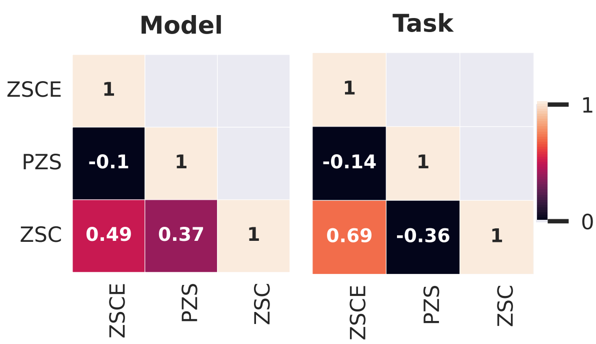

Figure 15: The left heatmap shows the correlation of the model rankings when changing the *base prompt*. The right heatmap shows the correlation of the task rankings when changing the *base prompt*. That means, how stable is the performance of all models across tasks, if the base prompt is changed. For both the model and for the task ranking, the change between Zero-Shot-CoT and Zero-Shot-CoT-EM keeps the ranking stable.

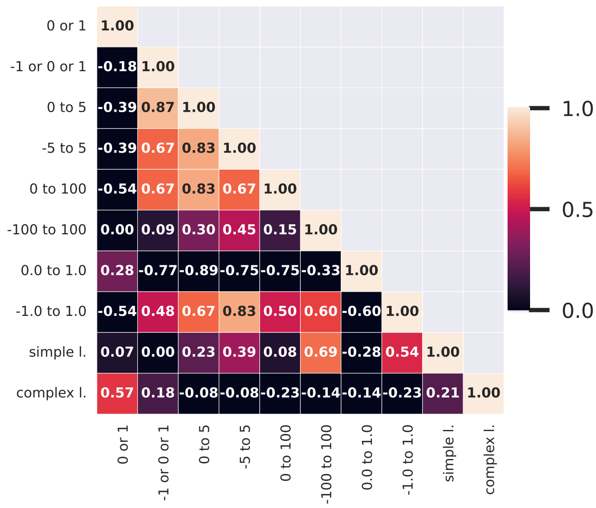

Figure 16: Correlation of the task rankings when changing the *format requirement*. That means, how stable is the performance of all models across tasks, if the format requirement is changed. Here, the stability when changing between format requirements is mixed. For some changes, like “0 to 5” and “-5 to 5” the ranking is very stable. For other changes, the ranking can change randomly or even be strongly negatively correlated. This means that considering all tested prompts (also weak performing ones) and models, their average correlation on task X might be the highest for format requirement 1 and the lowest for format requirement 2.