Diverse and High-Quality Text Generation Assisted by Large Language Models (Supplementary Material)

Thai Huy Nguyen [a,b], Ngoc Hoang Luong [a,b,*]

[a] *University of Information Technology, Ho Chi Minh City, Vietnam*
[b] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

The capability of many large pre-trained language models to generate natural-sounding, high-quality texts enables them to assist humans across various domains. However, the demand to generate a set of diverse, yet qualitative responses that satisfy different human preferences is often overlooked. Recently, Quality-Diversity (QD) algorithms have emerged as a promising approach for addressing these requirements in generation tasks, yet their applications in natural language generation remain under-explored. In this paper, we employ MAP-Elites, a conventional QD algorithm, to search for a wide range of responses with high-performing scores in different niches of hand-crafted features representing angles of human preferences. We adopt QD through AI Feedback (QDAIF) which uses large language models to modify (mutate) previously found texts into new ones and evaluate the value of such features and generation quality. However, QDAIF only focuses on creative writing tasks where there is no explicit task objective, unlike input instance-based tasks (e.g., summarizing a paragraph). We, thus, adjust the original pipeline to be more flexible on these tasks and further enhance the mutation procedure by introducing an assigned diversity mechanism which actively steers mutation towards unexplored preferences. Experiments show that our method successfully discovers diverse and high-quality texts both qualitatively and quantitatively in distinct categories of six natural language tasks including text continuation, abstractive summarization, data to text, generative commonsense, question answering and chitchat dialogue.

Appendix A. General Hyperparameters Details

Appendix A.1. Language Model Hyperparameters

Table A.1: Language model general hyperparameters used for the emitter and evaluator

Hyperparameters	Values
Top p	0.95
Temperature	1.1
Max generation length	2048
Sample token	True

Appendix A.2. Quality-Diversity and Genetic Algorithm hyperparameters

Table A.2: QD and GA search default hyperparameters used for all text generation tasks

Hyperparameters	Values
Total generations	50
Number of initial generations	10
Grid size	3
Number of niches (bins)	9
Batch size	1
Number of test instances	50 (`data2text`: 10)
Pool size (GA)	9

Table A.3: Ranges of descriptor for all tasks

Descriptors	Ranges of value
`length`	[0, 100] (`textsum`: [0, 15])
`formality`	[0, 10]
`maturity`	[0, 100]

Appendix B. Details of Tasks and Datasets

In the original GRUE benchmark [8], seven core tasks were proposed to evaluate reinforcement learning-based post-training methods for natural language understanding and generation. We adopt six of these tasks, excluding the translation task, for evaluating our approach. The translation task is omitted due to its requirement for high-fidelity outputs, which limits the possibility of generating diverse variations aligned with human preference descriptors. For the selected tasks, we retain the original datasets and adopt relevant evaluation metrics, while making several adjustments to the task objectives to better align with our quality-diversity framework. The modifications are detailed as follows:

- **Abstractive Summarization (`textsum`).** This task assesses a language model's capability to distill and rephrase the main ideas of a long input text into a concise summary,

*Corresponding author
Email addresses: `huynt.19@grad.uit.edu.vn` (Thai Huy Nguyen), `hoangln@uit.edu.vn` (Ngoc Hoang Luong)

reflecting the key content without directly copying. We use the CNN/Daily Mail dataset [2], which comprises news articles paired with human-written summaries. Each instance consists of a source paragraph and a reference summary. The task demands both content fidelity and abstraction, requiring the model to understand the full context and express it succinctly.

- **Text Continuation (`textcon`).** In this task, the model is prompted with the beginning of a text and must generate a plausible continuation. The focus is on coherence, fluency, and stylistic consistency. We construct this task using the IMDB dataset [6], which contains movie reviews. Each review is split into two parts: the first 15% of the original review is provided as input, while the remaining 85% serves as the reference. This setup allows us to evaluate the model's ability to generate text that is natural and consistent with the initial context.

- **Data-to-Text Generation (`data2text`).** This task evaluates the model's ability to interpret structured data and verbalize it in fluent natural language. We use the ToTTo dataset [7], which contains Wikipedia tables annotated with highlighted cells. Each instance provides a table, a subset of cells selected as relevant, and a reference sentence describing the highlighted content. The model must synthesize a coherent sentence that accurately reflects the underlying data while maintaining grammatical correctness and factual consistency.

- **Commonsense Text Generation (`commongen`).** This task measures the model's capacity to generate logically sound and contextually grounded text based on a set of everyday concepts. The CommonGEN dataset [5] supplies a list of 3–5 concept words for each instance, which the model must incorporate into a short passage. The generated text is evaluated based on commonsense plausibility, fluency, and its effective integration of all given concepts. This task challenges the model to express everyday reasoning in natural language.

- **Question Answering (`qa`).** In this task, the model is given a passage followed by a question and is required to generate an answer grounded in the input. We adopt the NarrativeQA dataset [3], where each example includes a narrative paragraph and a related question, along with a reference answer for evaluation. This setup tests the model's reading comprehension and ability to infer and extract relevant information for answering specific questions.

- **Chitchat Dialogue (`dialog`).** This task evaluates the model's performance in open-domain, multi-turn conversations. Given a dialogue history, the model must generate a contextually appropriate and coherent next utterance. We use the DailyDialog dataset [4], which features high-quality, manually annotated dialogues covering various topics. Each sample includes several conversational turns, and the model is required to continue the dialogue in a way that maintains consistency and conversational relevance.

Appendix C. Evaluator Details

We first present the pseudocode for the language model-based text evaluator in Algorithm 1. Additionally, Table C.4 details the perturbation process applied to the evaluation token in order to derive the set of selective tokens.

Algorithm 1: LM-based Text Evaluator

Initialize: set of perturbations $\{\textsc{Perturb}(text, i)\}_{i=1}^{N}$, a positive token pos, and a negative token neg, evaluation prompt $prompt_{eval}$
$logits \leftarrow \text{LM}(text, prompt_{eval})$
$pos\_logits \leftarrow \emptyset$
$neg\_logits \leftarrow \emptyset$
for $i = 1$ *to* N **do**
  $pos_i \leftarrow \textsc{Perturb}(pos, i)$
  $neg_i \leftarrow \textsc{Perturb}(neg, i)$
  $pos\_logits \leftarrow pos\_logits \cup logits[pos_i]$
  $neg\_logits \leftarrow neg\_logits \cup logits[neg_i]$
$Q(\text{text}) = \frac{\sum_{i=1}^{N} p^{(i)}_{\text{positive}}(\text{text})}{\sum_{i=1}^{N} [p^{(i)}_{\text{positive}}(\text{text}) + p^{(i)}_{\text{negative}}(\text{text})]}$

Table C.4: All the perturbation we apply on the selective token for quality and diversity assessment. We provide an example of the word 'yes' and its perturbations

Perturbation	Examples
Normal	'yes'
Spacing	' yes'
Capitalized First Character	'Yes'
Spacing & Capitalized First Character	' Yes'
All Capitalization	'YES'
Spacing & All Capitalization	' YES'

Appendix D. Emitter Details

In the emitter module, we prompt the language model to rewrite a given text (specifically, an offspring randomly sampled from the archive) according to a target set of human-aligned preferences. After selecting a niche, we randomly sample a descriptor vector located within that niche. However, directly conditioning the model on raw descriptor values for abstract preferences (e.g., "rewrite the text to be 10% the length of the original" or "target a formality score of 7") often provides limited guidance to the language model. To address this, we apply a transformation step to convert raw descriptor values into more interpretable prompt components.

The transformed values are either scalar quantities (e.g., for `length`) or discrete tokens (e.g., for `formality` and

`maturity`) that better communicate the intended preference. The transformation strategy is as follows:

`length`: For the `textsum` task, the length ratio is converted back to an estimated word count, which is used in the prompt. For other tasks, the raw value is retained.

`formality`: We define three categorical tokens: `informal`, `casual`, and `formal`, representing equal intervals along the formality axis. Given a descriptor value, we map it to the corresponding token and use it in the prompt to indicate the desired tone.

`maturity`: Similarly, we define three maturity-level tokens: `kids`, `teenagers`, and `adults`, that segment the maturity axis into equal thirds. The descriptor value is mapped to one of these tokens, which serves as the target audience specification in the prompt.

We note that in our framework, the evaluator operates using only the extreme ends of each descriptor spectrum, as this is sufficient to produce interpretable scores. Consequently, only two tokens are required to compute quality scores or descriptor values. In contrast, the emitter employs a finer-grained categorical design (three tokens in our implementation) to enable more precise control during generation and achieve broader coverage of the descriptor space.

Algorithm 2 provides the pseudocode of the assigned diversity mechanism in the emitter.

Algorithm 2: LLM-based Emitter with Assigned Diversity

Input: current archive $\mathcal{A}$, parent texts P
$sample\_prob \leftarrow \frac{1[s_i=\emptyset]}{\sum_{s\in\mathcal{A}} 1[s_i=\emptyset]}$
$target\_d \leftarrow$ SAMPLE($sample\_prob$)
$target\_prompt \leftarrow$ PROMPT_TRANSFORM($target\_d$)
$o \leftarrow$ GENERATE($p, target\_prompt$)
return o

Appendix E. Prompt Templates

In this section, we provide the prompts we used for the evaluator and the emitter in every task.

Appendix E.1. Abstractive Summarization

Quality evaluator prompts

Original paragraph: `input text`
Summary: `offspring text`
Does this summary capture the main contexts of the original paragraph? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What tone is this paragraph closest to from the following list: ['formal', 'informal']

Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a summary chatbot. Only write the summarization text.

Main prompt.
Original paragraph:
`input text`
Write an abstractive summary of the above paragraph for a `target age` with a `target formality` tone in exactly `target length` words. Only write the summarization text.

Diversity emitter

System prompt.
You are a summary chatbot. Only write the summarization text.

Main prompt.
Text:
`parent text`
Translate this summarization text into another summarization text for a `target age` with a `target formality` tone in exactly `target length` words but do not change the abstract information. Only write the summarization text.

Quality emitter

System prompt.
You are a summary chatbot. Only write the summarization text.

Main prompt.
Original paragraph:
`input text`
This is a summary:
`parent text`
Rewrite this summary into another summary that captures more abstract information from the original text without changing features. Only write the summarization text.

Appendix E.2. Text Continuation

Quality evaluator prompts

Unfinished paragraph: `input text`
Continuation: `offspring text`
Is this text a good continuation for the unfinished paragraph? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What kind of language is this paragraph closest to from the following list: ['formal', 'informal']

Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a chatbot who continues an unfinished paragraph. Only write the continuation text.

Main prompt.
Unfinished paragraph:
`input text`
Write a continuation of the above unfinished paragraph for a `target age` with a `target formality` tone in exactly `target length` words. Only write the continuation text.

Diversity emitter

System prompt.
You are a chatbot who continues an unfinished paragraph. Only write the continuation text.

Main prompt.
Text:
`parent text`
Translate this continuation text into another continuation text for a `target age` with a `target formality` tone in exactly `target length` words but do not change the abstract information. Only write the continuation text.

Quality emitter

System prompt.
You are a chatbot who continues an unfinished paragraph. Only write the continuation text.

Main prompt.
Original paragraph:
`input text`
This is a continuation:
`parent text`
Rewrite this continuation text into another continuation text that sounds more natural without changing the tone and length. Only write the continuation text.

Appendix E.3. Data to Text

Quality evaluator prompts

Table: `input text`
Highlighted cells: `highlighted cells`
Description: `offspring text`
Does this description capture all information of the highlighted cells in the given table and sound good? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What kind of language is this paragraph closest to from the following list: ['formal', 'informal']
Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a chatbot that generates a description of highlighted cells from a provided table. Only write the description text.
Main prompt.
Table:
`input text`
Highlighted cells:
`highlighted cells`
Write a description of the highlighted cells from the provided table for a `target age` with a `target formality` tone in exactly `target length` words. Only write the description text.

Diversity emitter

System prompt.
You are a chatbot that generates a description of highlighted cells from a provided table. Only write the description text.
Main prompt.
`parent text`
Translate this description text into another description text for a `target age` with a `target formality` tone in exactly `target length` words but do not change the abstract information. Only write the description text.

Quality emitter

System prompt.
You are a chatbot that generates a description of highlighted cells from a provided table. Only write the description text.
Main prompt.
Table:
`input text`
Highlighted cells:
`highlighted cells`
Description:
`parent text`
Rewrite this description into another description that sounds more natural and captures more information about the highlighted cells from the given table without changing the tone and length. Only write the description text.

Appendix E.4. Commonsense Generation

Quality evaluator prompts

Text: `offspring text`
Does this text contain all the words: `concepts` and sound good? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What kind of language is this paragraph closest to from the following list: ['formal', 'informal']
Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a chatbot who generates text from provided words. Only write the generated text.
Main prompt.
Provided words:
`concepts`
Write a paragraph containing all the provided words for a `target age` with a `target formality` tone in exactly `target length` words. Only write the generated text.

Diversity emitter

System prompt.
You are a chatbot who generates text from provided words. Only write the generated text.
Main prompt.
`parent text`
Translate this text into another text for a `target age` with a `target formality` tone in exactly `target length` words but still contains these words `concepts`. Only write the generated text.

Quality emitter

System prompt.
You are a chatbot who generates text from provided words. Only write the generated text.
Main prompt.
Table:
`input text`
Highlighted cells:
`highlighted cells`
Description:
`parent text`
Rewrite this text into another text that sound more natural and contains these words: `concepts` without changing the tone and length. Only write the generated text.

Appendix E.5. Question Answering

Quality evaluator prompts

Paragraph: `input text`
Question: `question`
Answer: `offspring text`
Does this answer correct and sound good to the question given the paragraph? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What kind of language is this paragraph closest to from the following list: ['formal', 'informal']

Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a chatbot who answer a question about a paragraph. Only write the answer text.

Main prompt.
Paragraph:
`input text`
Answer this question about the paragraph for a `target age` with a `target formality` tone in exactly `target length` words. Question: `question`. Only write the answer text.

Diversity emitter

System prompt.
You are a chatbot who answer a question about a paragraph. Only write the answer text.

Main prompt.
`parent text`
Translate the above answer into another answer for a `target age` with a `target formality` tone in exactly `target length` words. Only write the answer text.

Quality emitter

System prompt.
You are a chatbot who answer a question about a paragraph. Only write the answer text.

Main prompt.
Paragraph: `input text`
Question: `question`
Answer: `parent text`
Rewrite this answer into another more correct answer that sounds better and more natural without changing the tone and length given the paragraph. Only write the answer text.

Appendix E.6. Chitchat Dialogue

Quality evaluator prompts

Dialogue history: `original text`
Response: `offspring text`
Does this response sound good and natural for this dialogue history? Answer yes or no.

Descriptor evaluator prompts

Formality feedback.
`offspring text`
What kind of language is this paragraph closest to from the following list: ['formal', 'informal']

Maturity feedback.
`offspring text`
What audience is this paragraph more suitable for from the following list: ['adult', 'child']

Initialization

System prompt.
You are a chatbot who responses based on the dialogue history. Only write the response text.

Main prompt.
Continue this dialogue. for a `target age` with a `target formality` tone in exactly `target length` words. Question: `question`. Only write the response text.
Dialogue history:
`input text`

Diversity emitter

System prompt.
You are a chatbot who responses based on the dialogue history. Only write the response text.

Main prompt.
`parent text`
Translate the above response into another response for a `target age` with a `target formality` tone in exactly `target length` words. Only write the response text.

Quality emitter

System prompt.
You are a chatbot who responses based on the dialogue history. Only write the response text.

Main prompt.
Dialogue history: `original text`
Text: `parent text`
Rewrite this response into another response that sound better and more natural for the dialogue. Only write the response text.

Appendix F. Evaluator Reliability

While prior work [1] has shown that LLM-based evaluations exhibit strong correlation with human judgments on instance-free tasks (e.g., creative writing), it remains unclear whether this reliability also holds for instance-based natural language generation tasks, where outputs depend on specific inputs. To address this question, we examine the reliability of LLM-based evaluation for descriptor values assessments. In particular, we measure the Spearman correlation of descriptor values between two LLM evaluators, `Llama-3.2-3B-Instruct` and `Gemini-2.0-Flash`, to assess the consistency of their judgments. We first randomly sample 50 texts generated by `Llama-3.2-3B-Instruct` for each descriptor pair and task, and then compute the correlation over the aggregated set of these samples.

Table F.5: Spearman correlation of two descriptors between `Llama-3.2-3B-Instruct` and `Gemini-2.0-Flash`

Descriptors	Correlation values
`Formality`	0.77
`Maturity`	0.51

We report the correlation results in Table F.5. The results show that both `Formality` and `Maturity` exhibit strong positive correlations between the two models, with `Formality` showing a higher degree of agreement. This indicates that the two LLMs are largely consistent in their evaluations of descriptor values, demonstrating the reliability of using LLMs as evaluators in this setting.

Appendix G. Case Study

To demonstrate the effectiveness of our method in diversifying text generation under real-world conditions, we apply it to the task of summarizing a full-length scientific paper ("Attention Is All You Need" [9]) according to two user-defined preferences: `length` and `maturity`. The paper contains approximately 50,000 words, presenting a realistic challenge that reflects the scale and complexity of real-world applications. We conduct this experiment using the `Gemini-2.0-Flash` model. The descriptors are bounded within $[0, 5]$ for the ratio of generated text to the entire document and $[0, 100]$ for `length` and `maturity`, respectively. The resulting diverse summaries produced by our method, retrieved from the final archive, are presented below.

We observe that our method successfully explores the entire descriptor space, demonstrating its effectiveness in generating diverse summaries aligned with human-defined preferences. The resulting texts vary substantially in both length and tone, reflecting distinct communicative styles suitable for different audiences. For example, when targeting a lower `maturity` level, the model produces concise summaries written in a friendly and accessible tone, using simplified vocabulary and positive phrasing to make the Transformer architecture easier to grasp for younger readers. In contrast, for higher `maturity` levels, the summaries become more detailed and formal, preserving technical terminology and adopting a serious, academic tone appropriate for expert readers. These results highlight the flexibility of our approach, showing that it can adapt linguistic style, content depth, and narrative tone according to explicitly specified human preferences, thus supporting more personalized and context-aware text generation in real-world scenarios.

Example texts

Text: Imagine a super-smart robot translator! It's called the Transformer and uses "attention" to understand sentences quickly, better than old methods. It's amazing at translating languages like English to German! Plus, it's great at other tasks too, like understanding sentence structure. Scientists are making it even better to work with pictures and sounds too!
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 1.1613 / 2.0 / 30.0

Text: Imagine robots translating languages! Our robot, the Transformer, uses special "attention" to understand sentences without memorizing word orders like a song. It's super fast, learning to translate better than other robots in just one day! We even taught it to understand complicated grammar. Robots are getting smarter!
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 1.0298 / 2.0 / 35.0

Text: This paper introduces the Transformer, a novel neural network architecture for sequence transduction tasks like machine translation. Unlike traditional recurrent models, the Transformer relies entirely on attention mechanisms, enabling parallel processing and faster training. It achieves state-of-the-art results on English-to-German and English-to-French translation tasks, surpassing previous models with significantly lower training costs. The Transformer's architecture comprises stacked self-attention and fully connected layers, replacing recurrent layers. Experiments demonstrate its effectiveness and generalizability, extending to English constituency parsing. The authors propose future research directions, including applying the Transformer to other modalities and reducing sequential generation constraints.
Estimated Quality: 1.0
Estimated descriptor values (l/f/m): 2.0596 / 8.0 / 75.0

Text: Imagine a super smart robot that's awesome at understanding different languages! This robot, called a Transformer, learns faster than other robots because it uses "attention".
Instead of reading a sentence word by word, like reading slowly, it's more like skimming a book, picking out the important parts to quickly understand the whole story!
This way, it translates languages super fast and accurately. When tested with the trickiest language puzzles, it aced them!
The Transformer isn't just good at languages. It can also read long stories and make short summaries, just like a super-smart book reporter! Scientists believe that "attention" might help robots understand pictures, sounds, and videos, making them super smart at understanding the world.
The Transformer is a big leap forward in making robots smarter at learning. So many new AI systems will use the method of making translations faster, it will be super helpful in the future!
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 3.2647 / 3.0 / 30.0

Text: Okay, so picture this: a next-level translation robot called the Transformer! Forget the old-school robots that stumble through words one by one; this one's got a secret weapon: "attention."
Think of it as super-focus, pinpointing the most important bits of a sentence instantly. This "attention" thing lets it gulp down entire sentences at once, unlike those slow-reading "recurrent" bots.
Because it's so speedy, the Transformer learns languages ridiculously fast – think shortcut to language mastery! And guess what? It totally crushes the old champion translators in terms of accuracy.
But it's not just about translations, this robot is brilliant at dissecting sentences and understanding grammar. It's like a language Swiss Army knife, doing all sorts of cool stuff with words. The creators are buzzing with excitement about all the awesome possibilities and they're eager to see where this technology leads us.
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 3.0894 / 3.0 / 35.0

Text: The Transformer, a novel sequence transduction model, revolutionizes neural network architectures by eschewing recurrence in favor of multi-headed self-attention. This design enables superior parallelization, significantly accelerating training compared to recurrent and convolutional approaches.
In machine translation, the Transformer attains state-of-the-art results on WMT 2014 English-to-German and English-to-French tasks, surpassing previous models and ensembles in performance. It also exhibits strong generalization capabilities, achieving competitive results in English constituency parsing despite minimal task-specific tuning, outperforming RNN sequence-to-sequence models and the BerkeleyParser. The Transformer's success underscores the potential of attention-based models and opens avenues for future research. This includes expanding its application to diverse modalities like images, audio, and video, while exploring local attention mechanisms for efficient handling of large inputs and outputs. Further research will also focus on reducing sequential generation processes. This work is poised to shape the future of sequence modeling and transduction, with implications across various domains.
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 3.2428 / 8.0 / 85.0

Example texts

Text: Imagine you have a cool robot that can translate languages! This robot, called the Transformer, is super smart because it uses something called "attention." Attention is like focusing really hard on the important parts of a sentence to understand it better.
Usually, robots that translate languages use "recurrent" stuff, which means they have to read one word at a time, like a slow reader. But the Transformer is different! It can read all the words at once, like a super-fast scanner, because it uses attention instead of recurrent stuff.
Because the Transformer can read all the words at once, it can learn to translate languages much faster than other robots. It's like the Transformer is taking a shortcut to become a super translator!
And guess what? The Transformer is so good at translating that it's even better than the robots that used to be the best! It's like the Transformer is the new champion translator in the world.
But that's not all! The Transformer is also good at other things, like understanding how sentences are put together. It's like the Transformer is a super smart robot that can do lots of cool stuff with language.
The people who made the Transformer are really excited about it, and they want to use it to do even more amazing things in the future!
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 4.8422 / 1.0 / 25.0

Text: Hey, so these researchers came up with a new type of AI called the Transformer, which is super good at understanding and translating languages. Usually, AI models that do this use "recurrent layers," which are like thinking step-by-step.
But the Transformer does things differently. It uses "attention," which is like focusing on the most important parts of the sentence all at once. Think of it like reading a book by skimming to get all the points. Because it works this way, the Transformer can process stuff much faster and can understand and translate the text much more efficiently.
They tested the Transformer by giving it the most difficult translation tests, and it rocked them! It was able to do a way better job than all the other AIs.
What's cool is that this method of making translations faster can also be used for other types of AI. For example, these researchers have shown that it can be used in AI that reads and makes summaries of text. They think that this type of thinking can be helpful in other AI jobs, such as in understanding images, audio and video.
All in all, the Transformer is a big step in making AI better at understanding the world around us. It will surely be helpful in the AI field!
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 4.7546 / 3.0 / 50.0

Text: This paper introduces the Transformer, a novel neural network architecture designed for sequence transduction tasks like machine translation. Unlike traditional models that rely on recurrent or convolutional layers, the Transformer leverages attention mechanisms to model relationships between input and output elements. This allows for parallel processing of sequences, significantly reducing training time. The Transformer achieves state-of-the-art results on English-to-German and English-to-French translation benchmarks, surpassing previous models with substantially lower training costs.
The architecture consists of an encoder and decoder, both built from stacked self-attention and fully connected layers. Multi-head attention is a key component, enabling the model to capture diverse relationships within the data. Positional encoding is used to incorporate sequence order information. The authors compare self-attention to recurrent and convolutional layers, highlighting its advantages in computational complexity, parallelization, and path length for long-range dependencies.
Experiments demonstrate the effectiveness of various Transformer configurations, with larger models and dropout regularization leading to improved performance. Ablation studies explore the impact of different architectural choices, such as the number of attention heads and key dimensions. The Transformer's ability to generalize is showcased through experiments on English constituency parsing, achieving competitive results despite minimal task-specific tuning. This work opens new avenues for attention-based models in sequence transduction and beyond, paving the way for future research on handling diverse data modalities and further reducing sequential computation.
Estimated Quality: 0.9
Estimated descriptor values (l/f/m): 4.8422 / 8.0 / 85.0

Appendix H. Text Examples

To provide a more comprehensive qualitative analysis of our method's performance across different tasks, we include illustrative examples drawn from each evaluation domain. Specifically, we present two representative instances per task, each comprising the original input text alongside the human-annotated reference. For each instance, we showcase three solution texts retrieved from the archive constructed by the QD-diversity framework. These examples were selected to highlight the variation in outputs that satisfy distinct preferences while maintaining acceptable task performance. For each generated text, we report both the estimated evaluation metrics obtained from the language model and the ground-truth quality metrics computed using standard evaluation criteria. The language model-based evaluation includes the scalar quality score and three behavioral descriptors: `length` (l), which captures the verbosity or conciseness of the output; `formality` (f), which reflects the stylistic tone along the formal-informal axis; and `maturity` (m), which estimates the intended age-appropriateness of the language (e.g., child-oriented vs. adult-oriented content).

Appendix H.1. Abstractive Summarization

Sample 1

Sample #222212 in the CNN/Daily Mail dataset [2]

Example texts

Text: A family keeps their 22-pound, aggressive Himalayan cat, Lux, despite incident.
Estimated Quality: 0.9681
Estimated descriptor values (l/f/m): 2.5114 / 5.4505 / 85.1322
True quality: 'ROUGE-1': 0.1538, 'ROUGE-2': 0.08, 'ROUGE-L': 0.1154, 'ROUGE-L Sum': 0.1538, 'METEOR': 0.07, 'CHRF': 13.2443, 'BLEU': 0.0265, 'SacreBLEU': 0.9281, 'BERTScore': 0.8592, 'Perplexity': 221.4089

Text: A 22-pound feline, Lux, precipitated a hazardous situation by attacking his owners and a seven-month-old infant, resulting in confinement to their bedroom. Despite prior history of violence, the owners are opting for therapy over relinquishment, despite receiving adoption proposals. Lux will receive medical attention and therapy to address his aggression, and the family intends to retain his care.
Estimated Quality: 0.9567
Estimated descriptor values (l/f/m): 13.242 / 8.9385 / 93.605
True quality: 'ROUGE-1': 0.2772, 'ROUGE-2': 0.0606, 'ROUGE-L': 0.1584, 'ROUGE-L Sum': 0.2376, 'METEOR': 0.2078, 'CHRF': 28.9475, 'BLEU': 0.0282, 'SacreBLEU': 0.9947, 'BERTScore': 0.8531, 'Perplexity': 55.6812

Text: There's a super fierce cat named Lux that lives with a family in Portland. He's really mean and even bit their little baby, which was really scary. Even though that was a bad thing that happened, the family still loves Lux and wants to help him feel better. They've been taking extra special care of him and trying to make him less grumpy. Some people want to take Lux to live with a new family who can give him a quieter life, but the family wants to keep him home where he can stay happy and safe.
Estimated Quality: 0.956
Estimated descriptor values (l/f/m): 22.1461 / 1.0659 / 35.3111
True quality: 'ROUGE-1': 0.1857, 'ROUGE-2': 0.0145, 'ROUGE-L': 0.1286, 'ROUGE-L Sum': 0.1286, 'METEOR': 0.2, 'CHRF': 25.9232, 'BLEU': 0.0188, 'SacreBLEU': 0.6608, 'BERTScore': 0.8364, 'Perplexity': 22.9426

Sample 2

Sample #211243 in the CNN/Daily Mail dataset [2]

Example texts

Text: “Hey kids, there’s some not-so-good news about food prices. In the next 5 years, they’re going to go way up! That means the money we spend on food will cost almost an extra £850 each year, not just for special things like meat and yummy bread, but for everyday stuff like milk too. It’s because there’s not enough food and people want more meat and grains. To make things worse, some stuff we buy already costs a lot, so we can’t just get away with not spending more money on food.”
Estimated Quality: 0.9326
Estimated descriptor values (l/f/m): 25.1381 / 4.2344 / 32.1838
True quality: ’ROUGE-1’: 0.2105, ’ROUGE-2’: 0.0533, ’ROUGE-L’: 0.1316, ’ROUGE-L Sum’: 0.1579, ’METEOR’: 0.213, ’CHRF’: 27.2948, ’BLEU’: 0.0269, ’SacreBLEU’: 1.4404, ’BERTScore’: 0.851, ’Perplexity’: 20.3346

Text: Imagine all your favorite foods getting really expensive - globally, prices could rise by 20% in five years, making it harder for people to afford food and buying things they like.
Estimated Quality: 0.9599
Estimated descriptor values (l/f/m): 8.5635 / 6.9075 / 86.2949
True quality: ’ROUGE-1’: 0.1905, ’ROUGE-2’: 0.0, ’ROUGE-L’: 0.119, ’ROUGE-L Sum’: 0.1667, ’METEOR’: 0.1077, ’CHRF’: 20.2394, ’BLEU’: 0.026, ’SacreBLEU’: 0.9205, ’BERTScore’: 0.865, ’Perplexity’: 34.0856

Text: A global food shortage is expected to cause UK food prices to soar by nearly 20% over five years, increasing average household food bills by £850 and the annual food bill by £19.6 billion by 2018 due to rising demand and erratic weather.
Estimated Quality: 0.9617
Estimated descriptor values (l/f/m): 11.8785 / 9.7638 / 97.7952
True quality: ’ROUGE-1’: 0.3265, ’ROUGE-2’: 0.0625, ’ROUGE-L’: 0.2041, ’ROUGE-L Sum’: 0.2857, ’METEOR’: 0.2012, ’CHRF’: 26.7331, ’BLEU’: 0.0486, ’SacreBLEU’: 2.7103, ’BERTScore’: 0.8782, ’Perplexity’: 24.2854

Appendix H.2. Text Continuation

Sample 1

Sample #5036 in the IMDB dataset [6]

Example texts
Text: ...and meandering at times, losing focus on the main theme, causing the viewer to lose interest. The character development was satisfactory, but the dialogue at times felt forced and clichéd, detracting from the overall experience. Despite its flaws, the animation's visual elements and sound design were commendable, making it a satisfying, if not exceptional, watch for fans of the genre. **Estimated Quality:** 0.9231 **Estimated descriptor values (l/f/m):** 47.2441 / 6.583 / 67.6581 **True quality:** 'ROUGE-1': 0.1724, 'ROUGE-2': 0.0, 'ROUGE-L': 0.1264, 'ROUGE-L Sum': 0.1264, 'METEOR': 0.1162, 'CHRF': 23.3815, 'BLEU': 0.0198, 'SacreBLEU': 0.9935, 'BERTScore': 0.8412, 'Perplexity': 22.6549
Text: The movie wasn't exciting, and it didn't have any new surprises. It told the same old stories that we've heard before, and it didn't make us feel anything deeply. We didn't feel happy or sad, and we didn't think a lot about it when it was over. It was just a boring movie that didn't stay in our minds. **Estimated Quality:** 0.893 **Estimated descriptor values (l/f/m):** 46.4567 / 1.1488 / 37.377 **True quality:** 'ROUGE-1': 0.2584, 'ROUGE-2': 0.0227, 'ROUGE-L': 0.1348, 'ROUGE-L Sum': 0.1348, 'METEOR': 0.1549, 'CHRF': 19.8555, 'BLEU': 0.0213, 'SacreBLEU': 1.1541, 'BERTScore': 0.853, 'Perplexity': 13.3527
Text: ...longer than necessary, drawing out key plot points and frustrating the viewer's desire for resolution. The pacing, though generally well-balanced, suffered from excessive buildup and lackluster delivery of critical plot twists, detracting from the overall viewing experience and hindering the animation's overall impact and emotional resonance. This was a notable flaw. **Estimated Quality:** 0.8745 **Estimated descriptor values (l/f/m):** 40.1575 / 9.6858 / 52.4393 **True quality:** 'ROUGE-1': 0.1325, 'ROUGE-2': 0.0, 'ROUGE-L': 0.0964, 'ROUGE-L Sum': 0.0964, 'METEOR': 0.0957, 'CHRF': 22.0627, 'BLEU': 0.014, 'SacreBLEU': 0.4955, 'BERTScore': 0.8353, 'Perplexity': 45.9662

Sample 2

Sample #5680 in the IMDB dataset [6]

Example texts

Text: ...seemed to be going for - families with young kids, history buffs, and perhaps some people just looking for mindless entertainment. The result was a hodgepodge of random historical references, absurd set pieces, and an overbearing Nicholas Cage performance that somehow still managed to be both annoying and laughably bad at the same time. It's a shame, because the script actually had some decent ideas, but they were squandered in the film's hasty and incoherent pursuit of thrills and spills.
Estimated Quality: 0.9222
Estimated descriptor values (l/f/m): 19.9501 / 2.0483 / 49.4705
True quality: 'ROUGE-1': 0.1793, 'ROUGE-2': 0.0185, 'ROUGE-L': 0.0966, 'ROUGE-L Sum': 0.0966, 'METEOR': 0.067, 'CHRF': 12.8364, 'BLEU': 0.0011, 'SacreBLEU': 0.0737, 'BERTScore': 0.7985, 'Perplexity': 29.2722

Text: ...but I think the real problem is that it was definitely targeting the kind of super-enthusiastic kiddos who are eating up everything the Disney Channel serves up. The over-the-top treasure hunt, the wisecracking sidekick, and Nicholas Cage's grizzled but lovable everyman vibe were all right up the alley for the kind of movie that's trying to appeal to tweens and early teens. But when you try to pass it off as some kind of high-stakes adventure worthy of a true cinematic legend, it comes across as more like a cheesy Saturday morning cartoon than a bonafide movie experience.
Estimated Quality: 0.9494
Estimated descriptor values (l/f/m): 24.4389 / 2.4058 / 19.0316
True quality: 'ROUGE-1': 0.2227, 'ROUGE-2': 0.0219, 'ROUGE-L': 0.1092, 'ROUGE-L Sum': 0.1092, 'METEOR': 0.068, 'CHRF': 14.4376, 'BLEU': 0.0018, 'SacreBLEU': 0.1225, 'BERTScore': 0.7965, 'Perplexity': 26.2736

Text: ...was it trying to be a family-friendly, historical adventure film? Its tone is all over the place, veering wildly from slapstick comedy to overly dramatic action sequences. The plot itself is a jumbled mess of cryptic clues and over-the-top historical conspiracies. It's like the filmmakers took every historical reference from the 18th and 19th centuries, threw them into a blender, and hit puree. The result is a film that's equal parts confusing and cliched. And don't even get me started on the character of Ben Gates, played by Cage. He's a bumbling, egotistical treasure hunter who is supposed to be some kind of heroic figure, but comes across as more of a cartoon character. It's hard to take him seriously when he's shouting at people to "keep calm" and "follow the map" like some kind of goofy, 19th-century-era Indiana Jones knockoff. Overall, National Treasure is a film that's more annoying than exciting, and its overhyped reputation only adds to its frustrating experience.
Estimated Quality: 0.9459
Estimated descriptor values (l/f/m): 40.399 / 1.6484 / 31.2663
True quality: 'ROUGE-1': 0.2917, 'ROUGE-2': 0.0646, 'ROUGE-L': 0.1477, 'ROUGE-L Sum': 0.1477, 'METEOR': 0.1461, 'CHRF': 23.116, 'BLEU': 0.0143, 'SacreBLEU': 1.1225, 'BERTScore': 0.8085, 'Perplexity': 20.4459

Appendix H.3. Data to text

Sample 1

Input Text.
Sample #93766 in the ToTTo dataset [7]

Example texts

Text: Check out these sweet train details - here's the lowdown:
The highlighted cells are showin' off the train's identity number, aka the 'Car number', which is probably how you'll recognize each one.
So, the highlighted row under 'Car number' tells you when each train was officially handed over to its new company (that's the 'Delivery date' gig). And above that, you'll see the 'Manufacturer' brand name.
But that's not all - the 'Seating capacity' shows how many seats are on the entire train, and the 'Weight (t)' lets you know how much it all weighs, in tons.
Estimated Quality: 0.8808
Estimated descriptor values (l/f/m): 97.0 / 6.8094 / 58.9313
True quality: 'ROUGE-1': 0.087, 'ROUGE-2': 0.0, 'ROUGE-L': 0.0725, 'ROUGE-L Sum': 0.058, 'METEOR': 0.084, 'CHRF': 16.9588, 'BLEU': 0.0183, 'SacreBLEU': 0.8969, 'BERTScore': 0.8028, 'Perplexity': 27.5303

Text: There's a big bus with a special name 'KiHa 110-1' that was born in 1990. It can fit 52 kids inside! There are other buses just like it, and some were born in 1990 too, but made by different companies, just like how you might have friends with different moms and dads.
Estimated Quality: 0.7857
Estimated descriptor values (l/f/m): 52.0 / 0.5047 / 0.4208
True quality: 'ROUGE-1': 0.1978, 'ROUGE-2': 0.0449, 'ROUGE-L': 0.1538, 'ROUGE-L Sum': 0.1538, 'METEOR': 0.146, 'CHRF': 23.1751, 'BLEU': 0.042, 'SacreBLEU': 2.2501, 'BERTScore': 0.8392, 'Perplexity': 47.552

Text: A table containing car details such as delivery dates, manufacturers, and seating capacities is provided with selected data - Car numbers, deliveries from January 1990 and early March 1991, and manufacturers Fuji Heavy Industries and Niigata Tekkō. Seat capacities range, initial depot allocation of 52.
Estimated Quality: 0.9939
Estimated descriptor values (l/f/m): 45.0 / 8.5693 / 71.2505
True quality: 'ROUGE-1': 0.3704, 'ROUGE-2': 0.1013, 'ROUGE-L': 0.2716, 'ROUGE-L Sum': 0.2716, 'METEOR': 0.3307, 'CHRF': 35.3315, 'BLEU': 0.0824, 'SacreBLEU': 6.0141, 'BERTScore': 0.8748, 'Perplexity': 176.3786

Sample 2

Sample #79302 in the ToTTo dataset [7]

Example texts

Text: Here's a description of the highlighted cells from the table:
These highlighted cells show two of the teams that played in a soccer game. One team is Plovdiv, which was played on April 10, 2018. The score of the game was Plovdiv 2, Slavia Sofia 1.
Estimated Quality: 0.9832
Estimated descriptor values (l/f/m): 46.0 / 1.188 / 82.9749
True quality: 'ROUGE-1': 0.1154, 'ROUGE-2': 0.04, 'ROUGE-L': 0.1154, 'ROUGE-L Sum': 0.1154, 'METEOR': 0.2922, 'CHRF': 25.6939, 'BLEU': 0.0339, 'SacreBLEU': 1.6332, 'BERTScore': 0.8519, 'Perplexity': 31.9063

Text: Hey kids! This is where the soccer game results were stored. The game was between Plovdiv and Slavia Sofia, and it happened on April 10, 2018. Guess what? Plovdiv won with a score of 2-1!
Estimated Quality: 0.8985
Estimated descriptor values (l/f/m): 35.0 / 0.9467 / 4.3511
True quality: 'ROUGE-1': 0.1463, 'ROUGE-2': 0.0513, 'ROUGE-L': 0.1463, 'ROUGE-L Sum': 0.1463, 'METEOR': 0.3288, 'CHRF': 30.3747, 'BLEU': 0.0409, 'SacreBLEU': 1.9793, 'BERTScore': 0.8617, 'Perplexity': 28.844

Text: A dated sequence is observed in the highlighted cells. The top-left entry, 'April 10, 2018', establishes a precise temporal context. To the right, a two-point victory, '2-1', is recorded in a game-related metric. In the adjacent cell, 'Slavia Sofia', the victor is identified, suggesting a clear outcome of the engagement on the specified date.
Estimated Quality: 0.9491
Estimated descriptor values (l/f/m): 54.0 / 5.5665 / 97.3518
True quality: 'ROUGE-1': 0.0635, 'ROUGE-2': 0.0328, 'ROUGE-L': 0.0635, 'ROUGE-L Sum': 0.0635, 'METEOR': 0.08, 'CHRF': 16.8293, 'BLEU': 0.0174, 'SacreBLEU': 0.5239, 'BERTScore': 0.8218, 'Perplexity': 59.1452

Appendix H.4. Commonsense Generation

Sample 1

Sample #52155 in the CommonGEN dataset [5]

Example texts

Text: Effective navigation necessitates the deployment of a solid anchor. The ocean's formidable force demands a reliable means of stabilization. During maritime excursions, anchoring on the coast provides a secure foundation. By harnessing the anchor's efficacy, sailors can mitigate risks and enjoy a tranquil experience, underscoring the importance of this maritime tool.
Estimated Quality: 0.9508
Estimated descriptor values (l/f/m): 51.0 / 9.3236 / 96.0593
True quality: 'ROUGE-1': 0.1818, 'ROUGE-2': 0.0312, 'ROUGE-L': 0.1212, 'ROUGE-L Sum': 0.1212, 'METEOR': 0.1695, 'CHRF': 23.0124, 'BLEU': 0.0309, 'SacreBLEU': 1.4886, 'BERTScore': 0.8385, 'Perplexity': 52.7132

Text: The young adventurers anchored themselves safely on the coast as they braved the strong stormy force together together.
Estimated Quality: 0.9621
Estimated descriptor values (l/f/m): 18.0 / 2.2785 / 16.9979
True quality: 'ROUGE-1': 0.2581, 'ROUGE-2': 0.069, 'ROUGE-L': 0.1935, 'ROUGE-L Sum': 0.1935, 'METEOR': 0.1471, 'CHRF': 30.0514, 'BLEU': 0.0963, 'SacreBLEU': 4.815, 'BERTScore': 0.8681, 'Perplexity': 182.9795

Text: Imagine you're on a big beach with lots of ships sailing by. There's this super cool lighthouse that's like a big friendly anchor, helping everyone stay safe and on track. It shines its light really far out into the sea, like a powerful force that guides ships all the way along the coast. And just like that, the ships can navigate through the ocean with ease!
Estimated Quality: 0.9342
Estimated descriptor values (l/f/m): 66.0 / 2.3551 / 3.8676
True quality: 'ROUGE-1': 0.1463, 'ROUGE-2': 0.025, 'ROUGE-L': 0.0976, 'ROUGE-L Sum': 0.0976, 'METEOR': 0.1554, 'CHRF': 23.7998, 'BLEU': 0.0257, 'SacreBLEU': 1.2437, 'BERTScore': 0.8313, 'Perplexity': 32.1791

Sample 2

Sample #30350 in the CommonGEN dataset [5]

Example texts

Text: "Hey there, super cool kid! Do you know what it means when you really, really LOVE something? Like, something so much you just can't imagine playing with anything else! That's totally awesome, and it's really good that you have something that brings you so much joy. Just remember, it's also important to make friends and have fun with other people, and you can always play with your favorite toy, but also with your friends. You're an amazing kid, and it's okay to LOVE toys, but don't forget to love your friends and family too!"
Estimated Quality: 0.9659
Estimated descriptor values (l/f/m): 94.0 / 0.2039 / 0.4788
True quality: 'ROUGE-1': 0.0541, 'ROUGE-2': 0.0, 'ROUGE-L': 0.0541, 'ROUGE-L Sum': 0.0541, 'METEOR': 0.0922, 'CHRF': 10.5122, 'BLEU': 0.0128, 'SacreBLEU': 0.425, 'BERTScore': 0.8276, 'Perplexity': 15.5638

Text: That cherished childhood toy now serves as a poignant reminder of the love we must nurture in adulthood, trading reckless carefree youth days for mature responsibilities, one which will become a lasting foundation of a meaningful life, much like a kid's favorite toy can become a tangible symbol of enduring love, reminding us that, with age comes a deeper appreciation for life's value and an unwavering dedication to those we hold dear.
Estimated Quality: 0.924
Estimated descriptor values (l/f/m): 72.0 / 8.6639 / 88.2325
True quality: 'ROUGE-1': 0.119, 'ROUGE-2': 0.0488, 'ROUGE-L': 0.0952, 'ROUGE-L Sum': 0.0952, 'METEOR': 0.1955, 'CHRF': 12.7753, 'BLEU': 0.0217, 'SacreBLEU': 1.023, 'BERTScore': 0.8336, 'Perplexity': 47.8791

Text: Childhood memories of playtime with a favorite toy evoke enduring love as a kid grows.
Estimated Quality: 0.9339
Estimated descriptor values (l/f/m): 15.0 / 1.0761 / 45.1742
True quality: 'ROUGE-1': 0.4, 'ROUGE-2': 0.1739, 'ROUGE-L': 0.32, 'ROUGE-L Sum': 0.32, 'METEOR': 0.4444, 'CHRF': 26.1895, 'BLEU': 0.1514, 'SacreBLEU': 9.7824, 'BERTScore': 0.8745, 'Perplexity': 150.52

Appendix H.5. Question Answering

Sample 1

Sample #25343 in the NarrativeQA dataset [3]

Example texts

Text: Judy was forced to attend a boarding school by her father, Captain Woolcot, after she pulled off a prank at his military base that humiliated him.
Estimated Quality: 0.9252
Estimated descriptor values (l/f/m): 26.0 / 7.5009 / 35.504
True quality: 'ROUGE-1': 0.3902, 'ROUGE-2': 0.1538, 'ROUGE-L': 0.1951, 'ROUGE-L Sum': 0.1951, 'METEOR': 0.3849, 'CHRF': 33.7738, 'BLEU': 0.086, 'SacreBLEU': 4.7897, 'BERTScore': 0.8784, 'Perplexity': 57.9623

Text: She was subsequently enrolled in a boarding school as a consequence of a prank orchestrated by herself and her sibling Pip, which had humiliated her father in their military quarters.
Estimated Quality: 0.9312
Estimated descriptor values (l/f/m): 30.0 / 5.5816 / 32.5949
True quality: 'ROUGE-1': 0.3556, 'ROUGE-2': 0.186, 'ROUGE-L': 0.3556, 'ROUGE-L Sum': 0.3556, 'METEOR': 0.3186, 'CHRF': 33.0196, 'BLEU': 0.0823, 'SacreBLEU': 4.6399, 'BERTScore': 0.8787, 'Perplexity': 72.2684

Text: Judy's prank ruined her dad's rep, getting her barred from home.
Estimated Quality: 0.8792
Estimated descriptor values (l/f/m): 11.0 / 0.6247 / 42.0268
True quality: 'ROUGE-1': 0.1429, 'ROUGE-2': 0.0, 'ROUGE-L': 0.1429, 'ROUGE-L Sum': 0.1429, 'METEOR': 0.1333, 'CHRF': 17.5994, 'BLEU': 0.0835, 'SacreBLEU': 2.9887, 'BERTScore': 0.8872, 'Perplexity': 148.9594

Sample 2

Sample #2814 in the NarrativeQA dataset [3]

Example texts

Text: Grosse Pointe, Michigan.
Estimated Quality: 0.9615
Estimated descriptor values (l/f/m): 3.0 / 7.7296 / 91.2966
True quality: 'ROUGE-1': 0.5714, 'ROUGE-2': 0.4, 'ROUGE-L': 0.5714, 'ROUGE-L Sum': 0.5714, 'METEOR': 0.3, 'CHRF': 57.6372, 'BLEU': 0.3247, 'SacreBLEU': 14.0585, 'BERTScore': 0.9552, 'Perplexity': 21.9978

Text: Hey kiddo, so there's some dude named Martin and he's having a reunion with old friends from high school, and they're all going back to his hometown, which is called Grosse Pointe, Michigan!
Estimated Quality: 0.9469
Estimated descriptor values (l/f/m): 33.0 / 0.8561 / 45.4609
True quality: 'ROUGE-1': 0.1, 'ROUGE-2': 0.0526, 'ROUGE-L': 0.1, 'ROUGE-L Sum': 0.1, 'METEOR': 0.1744, 'CHRF': 33.2116, 'BLEU': 0.0377, 'SacreBLEU': 1.2754, 'BERTScore': 0.8538, 'Perplexity': 24.2199

Text: Martin's high school reunion is held in his hometown of Grosse Pointe, Michigan. He attends the reunion under pressure from his therapist and secretary to smooth things over with a client, and coincidentally meets there due to a booked contract. Martin reconnects with friends, makes amends with his past, and gets involved in a deadly situation. The reunion provides a pivotal moment in the story.
Estimated Quality: 0.9613
Estimated descriptor values (l/f/m): 65.0 / 7.2288 / 99.4425
True quality: 'ROUGE-1': 0.0571, 'ROUGE-2': 0.0294, 'ROUGE-L': 0.0571, 'ROUGE-L Sum': 0.0571, 'METEOR': 0.1261, 'CHRF': 21.3777, 'BLEU': 0.0195, 'SacreBLEU': 0.6509, 'BERTScore': 0.8514, 'Perplexity': 32.7694

Appendix H.6. Chitchat Dialogue

Sample 1

Sample #4879 in the DailyDialog dataset [4]

Example texts
Text: "Hey kiddo, to get to the post office, go down the street, pass the coffee shop, turn right when you see the third sign, and you'll find it with a big flag waving, take the stairs to the second floor." **Estimated Quality:** 0.9646 **Estimated descriptor values (l/f/m):** 40.0 / 0.9613 / 6.1349 **True quality:** 'ROUGE-1': 0.2623, 'ROUGE-2': 0.1017, 'ROUGE-L': 0.1639, 'ROUGE-L Sum': 0.1639, 'METEOR': 0.2503, 'CHRF': 30.4601, 'BLEU': 0.0412, 'SacreBLEU': 2.0367, 'BERTScore': 0.856, 'Perplexity': 22.3757
Text: It's right down the street from here, you can't miss it. Just walk straight down this road, pass the ice cream shop, and take the left turn onto Main St. You'll see it on your right, it's got a big sign that says "Post Office". Easy peasy! **Estimated Quality:** 0.9852 **Estimated descriptor values (l/f/m):** 47.0 / 0.3185 / 16.3866 **True quality:** 'ROUGE-1': 0.3099, 'ROUGE-2': 0.058, 'ROUGE-L': 0.169, 'ROUGE-L Sum': 0.169, 'METEOR': 0.3366, 'CHRF': 27.4303, 'BLEU': 0.0383, 'SacreBLEU': 1.9056, 'BERTScore': 0.862, 'Perplexity': 15.5054
Text: To ensure you reach the post office correctly, please proceed in the direction of the road, pass the ice cream establishment located on your left-hand side, and then make a left turn onto Main Street. The post office will be situated on your right-hand side, characterized by its clear and prominent signage. **Estimated Quality:** 0.9906 **Estimated descriptor values (l/f/m):** 52.0 / 7.1509 / 52.7757 **True quality:** 'ROUGE-1': 0.3784, 'ROUGE-2': 0.1667, 'ROUGE-L': 0.2973, 'ROUGE-L Sum': 0.2973, 'METEOR': 0.5418, 'CHRF': 38.1896, 'BLEU': 0.1157, 'SacreBLEU': 9.7916, 'BERTScore': 0.8711, 'Perplexity': 22.3402

Sample 2

Sample #1046 in the DailyDialog dataset [4]

Example texts

Text: Sounds good! Let's make it more attractive to sellers - how about we throw in some perks like fast payouts, easy listing, and access to a large customer base? We could also offer exclusive design options, competitive fees, and top-notch customer support. What do you think? Should we add any other features to sweeten the deal?
Estimated Quality: 0.9716
Estimated descriptor values (l/f/m): 56.0 / 5.7767 / 76.794
True quality: 'ROUGE-1': 0.129, 'ROUGE-2': 0.1, 'ROUGE-L': 0.129, 'ROUGE-L Sum': 0.129, 'METEOR': 0.3962, 'CHRF': 20.5775, 'BLEU': 0.0513, 'SacreBLEU': 3.4862, 'BERTScore': 0.8516, 'Perplexity': 24.5387

Text: "Let's start by gathering information about your current car and the market demand for it. We can also prepare a list of features and pricing to attract potential buyers. Additionally, we can advertise your car online and in local newspapers to increase visibility. With a little effort, we can sell your car quickly and at a good price, giving you an opportunity to break your habit of relying on it."
Estimated Quality: 0.9735
Estimated descriptor values (l/f/m): 70.0 / 5.2685 / 99.7518
True quality: 'ROUGE-1': 0.0789, 'ROUGE-2': 0.027, 'ROUGE-L': 0.0526, 'ROUGE-L Sum': 0.0526, 'METEOR': 0.2207, 'CHRF': 7.4258, 'BLEU': 0.0168, 'SacreBLEU': 0.5426, 'BERTScore': 0.8299, 'Perplexity': 20.1506

Text: Sell the car, get a bike, and ride downtown easily.
Estimated Quality: 0.9695
Estimated descriptor values (l/f/m): 10.0 / 0.5925 / 90.744
True quality: 'ROUGE-1': 0.0, 'ROUGE-2': 0.0, 'ROUGE-L': 0.0, 'ROUGE-L Sum': 0.0, 'METEOR': 0.0658, 'CHRF': 8.5373, 'BLEU': 0.0803, 'SacreBLEU': 0.0, 'BERTScore': 0.8381, 'Perplexity': 69.1158

Appendix I. Full Results

We present the complete results across all evaluation metrics, reported in two forms: Max Score and QD Score. Each table reports the mean and standard deviation computed over all test samples, for each task and method. The best-performing results are highlighted in bold, while results that are statistically significantly different from the best ($p < 0.05$) are shaded in gray.

Table I.6: The Coverage results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.38_{\pm 0.18}$	$0.86_{\pm 0.09}$	$0.71_{\pm 0.23}$	$0.38_{\pm 0.17}$	$0.59_{\pm 0.21}$	$0.59_{\pm 0.22}$
GA	$0.4_{\pm 0.14}$	$0.42_{\pm 0.15}$	$0.4_{\pm 0.13}$	$0.37_{\pm 0.13}$	$0.37_{\pm 0.15}$	$0.41_{\pm 0.13}$
QD init only	$0.74_{\pm 0.21}$	$0.84_{\pm 0.13}$	$0.9_{\pm 0.1}$	$0.78_{\pm 0.18}$	$0.88_{\pm 0.14}$	$0.86_{\pm 0.13}$
QD-quality	$0.65_{\pm 0.23}$	$0.72_{\pm 0.12}$	$0.82_{\pm 0.15}$	$0.71_{\pm 0.15}$	$0.84_{\pm 0.19}$	$0.84_{\pm 0.16}$
QD-diversity	$\mathbf{0.97}_{\pm 0.1}$	$\mathbf{1.0}_{\pm 0.0}$	$\mathbf{0.93}_{\pm 0.09}$	$\mathbf{0.97}_{\pm 0.06}$	$\mathbf{0.96}_{\pm 0.09}$	$\mathbf{0.94}_{\pm 0.08}$
			Length/Maturity			
LLM sampling	$0.28_{\pm 0.09}$	$0.58_{\pm 0.13}$	$0.7_{\pm 0.27}$	$0.32_{\pm 0.15}$	$0.5_{\pm 0.19}$	$0.44_{\pm 0.15}$
GA	$0.35_{\pm 0.1}$	$0.38_{\pm 0.11}$	$0.41_{\pm 0.14}$	$0.3_{\pm 0.12}$	$0.26_{\pm 0.11}$	$0.28_{\pm 0.1}$
QD init only	$0.73_{\pm 0.19}$	$0.7_{\pm 0.15}$	$0.95_{\pm 0.07}$	$0.64_{\pm 0.18}$	$0.87_{\pm 0.14}$	$0.68_{\pm 0.2}$
QD-quality	$0.52_{\pm 0.18}$	$0.58_{\pm 0.16}$	$0.85_{\pm 0.14}$	$0.55_{\pm 0.16}$	$0.71_{\pm 0.22}$	$0.58_{\pm 0.21}$
QD-diversity	$\mathbf{0.97}_{\pm 0.11}$	$\mathbf{0.97}_{\pm 0.09}$	$\mathbf{0.99}_{\pm 0.03}$	$\mathbf{0.99}_{\pm 0.05}$	$\mathbf{0.99}_{\pm 0.04}$	$\mathbf{0.98}_{\pm 0.05}$
			Formality/Maturity			
LLM sampling	$0.23_{\pm 0.13}$	$0.59_{\pm 0.14}$	$0.62_{\pm 0.2}$	$0.36_{\pm 0.14}$	$0.49_{\pm 0.21}$	$0.42_{\pm 0.14}$
GA	$0.24_{\pm 0.11}$	$0.28_{\pm 0.12}$	$0.32_{\pm 0.13}$	$0.29_{\pm 0.12}$	$0.29_{\pm 0.13}$	$0.28_{\pm 0.1}$
QD init only	$0.64_{\pm 0.15}$	$0.72_{\pm 0.16}$	$\mathbf{0.73}_{\pm 0.15}$	$0.47_{\pm 0.17}$	$0.66_{\pm 0.15}$	$0.52_{\pm 0.15}$
QD-quality	$0.46_{\pm 0.12}$	$0.63_{\pm 0.18}$	$0.69_{\pm 0.18}$	$0.48_{\pm 0.16}$	$0.63_{\pm 0.19}$	$0.49_{\pm 0.18}$
QD-diversity	$\mathbf{0.84}_{\pm 0.14}$	$\mathbf{0.89}_{\pm 0.11}$	$0.62_{\pm 0.14}$	$\mathbf{0.76}_{\pm 0.16}$	$\mathbf{0.73}_{\pm 0.15}$	$\mathbf{0.62}_{\pm 0.13}$

Table I.7: The QD-score results of multiple samples for each task in different pairs of preferences on LLAMA-3.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$3.16_{\pm 1.51}$	$6.8_{\pm 2.33}$	$5.97_{\pm 1.92}$	$3.27_{\pm 1.47}$	$4.67_{\pm 1.69}$	$5.17_{\pm 1.91}$
GA	$3.33_{\pm 1.23}$	$3.18_{\pm 1.56}$	$3.43_{\pm 1.07}$	$3.22_{\pm 1.1}$	$3.08_{\pm 1.27}$	$3.64_{\pm 1.12}$
QD init only	$6.16_{\pm 1.79}$	$6.76_{\pm 2.49}$	$7.6_{\pm 0.85}$	$6.71_{\pm 1.59}$	$7.02_{\pm 1.18}$	$7.6_{\pm 1.17}$
QD-quality	$5.4_{\pm 1.9}$	$5.18_{\pm 2.03}$	$6.9_{\pm 1.18}$	$5.49_{\pm 1.54}$	$6.75_{\pm 1.57}$	$7.39_{\pm 1.37}$
QD-diversity	$\mathbf{8.42}_{\pm 0.9}$	$\mathbf{7.67}_{\pm 2.48}$	$\mathbf{7.81}_{\pm 0.81}$	$\mathbf{8.22}_{\pm 0.58}$	$\mathbf{7.59}_{\pm 0.79}$	$\mathbf{8.23}_{\pm 0.76}$
			Length/Maturity			
LLM sampling	$2.29_{\pm 0.76}$	$4.58_{\pm 1.85}$	$5.83_{\pm 2.24}$	$2.79_{\pm 1.27}$	$3.99_{\pm 1.5}$	$3.91_{\pm 1.32}$
GA	$2.93_{\pm 0.83}$	$2.97_{\pm 1.33}$	$3.51_{\pm 1.21}$	$2.56_{\pm 1.08}$	$2.19_{\pm 0.94}$	$2.48_{\pm 0.88}$
QD init only	$6.02_{\pm 1.52}$	$5.61_{\pm 2.17}$	$7.93_{\pm 0.64}$	$5.5_{\pm 1.57}$	$6.86_{\pm 1.22}$	$6.02_{\pm 1.75}$
QD-quality	$4.29_{\pm 1.39}$	$4.4_{\pm 1.89}$	$7.07_{\pm 1.2}$	$4.22_{\pm 1.36}$	$5.69_{\pm 1.77}$	$5.16_{\pm 1.87}$
QD-diversity	$\mathbf{8.27}_{\pm 0.97}$	$\mathbf{7.19}_{\pm 2.45}$	$\mathbf{8.32}_{\pm 0.34}$	$\mathbf{8.24}_{\pm 0.66}$	$\mathbf{7.81}_{\pm 0.46}$	$\mathbf{8.58}_{\pm 0.46}$
			Formality/Maturity			
LLM sampling	$1.92_{\pm 1.08}$	$4.86_{\pm 1.91}$	$5.19_{\pm 1.68}$	$3.15_{\pm 1.24}$	$3.9_{\pm 1.67}$	$3.66_{\pm 1.24}$
GA	$2.03_{\pm 0.92}$	$2.24_{\pm 1.23}$	$2.75_{\pm 1.14}$	$2.56_{\pm 1.07}$	$2.43_{\pm 1.08}$	$2.46_{\pm 0.85}$
QD init only	$5.21_{\pm 1.17}$	$5.71_{\pm 2.24}$	$\mathbf{6.12}_{\pm 1.25}$	$4.02_{\pm 1.41}$	$5.29_{\pm 1.22}$	$4.56_{\pm 1.33}$
QD-quality	$3.78_{\pm 0.93}$	$5.06_{\pm 2.19}$	$5.83_{\pm 1.51}$	$3.72_{\pm 1.39}$	$5.1_{\pm 1.57}$	$4.29_{\pm 1.6}$
QD-diversity	$\mathbf{7.3}_{\pm 1.2}$	$\mathbf{7.16}_{\pm 1.81}$	$5.19_{\pm 1.11}$	$\mathbf{6.4}_{\pm 1.41}$	$\mathbf{5.92}_{\pm 1.24}$	$\mathbf{5.49}_{\pm 1.15}$

Table I.8: The Max fitness results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.93_{\pm 0.05}$	$\mathbf{0.9}_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.91_{\pm 0.04}$	$0.99_{\pm 0.01}$
GA	$0.94_{\pm 0.04}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$\mathbf{0.94}_{\pm 0.03}$	$\mathbf{0.99}_{\pm 0.01}$
QD init only	$0.94_{\pm 0.04}$	$0.9_{\pm 0.29}$	$0.97_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.01}$	$0.92_{\pm 0.03}$	$0.99_{\pm 0.01}$
QD-quality	$0.94_{\pm 0.04}$	$0.84_{\pm 0.28}$	$0.96_{\pm 0.01}$	$0.9_{\pm 0.08}$	$0.93_{\pm 0.03}$	$0.99_{\pm 0.0}$
QD-diversity	$\mathbf{0.98}_{\pm 0.01}$	$0.89_{\pm 0.29}$	$\mathbf{0.97}_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.93_{\pm 0.03}$	$0.99_{\pm 0.01}$
			Length/Maturity			
LLM sampling	$0.93_{\pm 0.05}$	$\mathbf{0.9}_{\pm 0.29}$	$0.95_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.91_{\pm 0.03}$	$0.99_{\pm 0.01}$
GA	$0.94_{\pm 0.04}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.02}$	$\mathbf{0.94}_{\pm 0.03}$	$\mathbf{0.99}_{\pm 0.01}$
QD init only	$0.94_{\pm 0.04}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.01}$	$0.92_{\pm 0.03}$	$0.99_{\pm 0.01}$
QD-quality	$0.94_{\pm 0.04}$	$0.84_{\pm 0.28}$	$0.95_{\pm 0.01}$	$0.9_{\pm 0.08}$	$0.93_{\pm 0.03}$	$0.99_{\pm 0.01}$
QD-diversity	$\mathbf{0.98}_{\pm 0.01}$	$0.89_{\pm 0.29}$	$\mathbf{0.96}_{\pm 0.01}$	$0.97_{\pm 0.02}$	$0.92_{\pm 0.03}$	$0.99_{\pm 0.01}$
			Formality/Maturity			
LLM sampling	$0.93_{\pm 0.05}$	$0.9_{\pm 0.29}$	$0.95_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.91_{\pm 0.03}$	$0.99_{\pm 0.01}$
GA	$0.94_{\pm 0.05}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$\mathbf{0.95}_{\pm 0.02}$	$\mathbf{0.99}_{\pm 0.0}$
QD init only	$0.93_{\pm 0.05}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.92_{\pm 0.03}$	$0.99_{\pm 0.01}$
QD-quality	$0.93_{\pm 0.05}$	$0.9_{\pm 0.29}$	$0.96_{\pm 0.01}$	$0.9_{\pm 0.08}$	$0.94_{\pm 0.03}$	$0.99_{\pm 0.01}$
QD-diversity	$\mathbf{0.98}_{\pm 0.01}$	$\mathbf{0.94}_{\pm 0.21}$	$\mathbf{0.96}_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.02}$	$0.94_{\pm 0.03}$	$0.99_{\pm 0.01}$

Table I.9: The Max BERTScore results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.87_{\pm 0.01}$	$0.84_{\pm 0.02}$	$0.87_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.04}$	$0.83_{\pm 0.01}$	$0.84_{\pm 0.02}$
GA	$0.86_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.87_{\pm 0.02}$	$0.87_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.83_{\pm 0.02}$
QD init only	$\mathbf{0.87}_{\pm 0.01}$	$\mathbf{0.85}_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.02}$	$0.89_{\pm 0.03}$	$\mathbf{0.83}_{\pm 0.01}$	$0.84_{\pm 0.02}$
QD-quality	$0.87_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.88_{\pm 0.02}$	$0.87_{\pm 0.03}$	$0.83_{\pm 0.01}$	$0.84_{\pm 0.02}$
QD-diversity	$0.87_{\pm 0.01}$	$0.85_{\pm 0.02}$	$0.88_{\pm 0.02}$	$0.88_{\pm 0.03}$	$0.83_{\pm 0.01}$	$\mathbf{0.84}_{\pm 0.02}$
			Length/Maturity			
LLM sampling	$0.87_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.88_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.84_{\pm 0.02}$
GA	$0.86_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.87_{\pm 0.02}$	$0.87_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.83_{\pm 0.02}$
QD init only	$\mathbf{0.87}_{\pm 0.02}$	$0.86_{\pm 0.02}$	$\mathbf{0.89}_{\pm 0.02}$	$0.88_{\pm 0.03}$	$\mathbf{0.83}_{\pm 0.01}$	$\mathbf{0.85}_{\pm 0.02}$
QD-quality	$0.86_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.88_{\pm 0.02}$	$0.87_{\pm 0.04}$	$0.83_{\pm 0.01}$	$0.84_{\pm 0.02}$
QD-diversity	$0.87_{\pm 0.02}$	$\mathbf{0.86}_{\pm 0.02}$	$0.88_{\pm 0.02}$	$0.88_{\pm 0.03}$	$0.83_{\pm 0.01}$	$0.84_{\pm 0.02}$
			Formality/Maturity			
LLM sampling	$\mathbf{0.87}_{\pm 0.01}$	$0.84_{\pm 0.03}$	$0.86_{\pm 0.01}$	$\mathbf{0.9}_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.84_{\pm 0.02}$
GA	$0.85_{\pm 0.02}$	$0.83_{\pm 0.02}$	$0.85_{\pm 0.01}$	$0.87_{\pm 0.04}$	$0.82_{\pm 0.01}$	$0.83_{\pm 0.02}$
QD init only	$0.86_{\pm 0.01}$	$0.84_{\pm 0.02}$	$\mathbf{0.86}_{\pm 0.01}$	$0.89_{\pm 0.03}$	$\mathbf{0.82}_{\pm 0.01}$	$\mathbf{0.84}_{\pm 0.02}$
QD-quality	$0.86_{\pm 0.01}$	$\mathbf{0.84}_{\pm 0.02}$	$0.86_{\pm 0.01}$	$0.86_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.84_{\pm 0.02}$
QD-diversity	$0.85_{\pm 0.02}$	$0.84_{\pm 0.02}$	$0.86_{\pm 0.01}$	$0.88_{\pm 0.03}$	$0.82_{\pm 0.01}$	$0.84_{\pm 0.02}$

Table I.10: The QD BERTScore results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$2.96_{\pm 1.37}$	$6.31_{\pm 0.69}$	$5.47_{\pm 1.76}$	$2.96_{\pm 1.31}$	$4.33_{\pm 1.56}$	$4.39_{\pm 1.62}$
GA	$3.01_{\pm 1.1}$	$3.1_{\pm 1.11}$	$3.09_{\pm 0.98}$	$2.84_{\pm 0.96}$	$2.73_{\pm 1.14}$	$3.05_{\pm 0.94}$
QD init only	$5.74_{\pm 1.64}$	$6.23_{\pm 1.01}$	$6.96_{\pm 0.8}$	$5.98_{\pm 1.42}$	$6.5_{\pm 1.03}$	$6.44_{\pm 1.01}$
QD-quality	$4.96_{\pm 1.72}$	$5.3_{\pm 0.93}$	$6.34_{\pm 1.14}$	$5.39_{\pm 1.19}$	$6.18_{\pm 1.4}$	$6.22_{\pm 1.15}$
QD-diversity	$\mathbf{7.43}_{\pm 0.79}$	$\mathbf{7.44}_{\pm 0.11}$	$\mathbf{7.2}_{\pm 0.72}$	$\mathbf{7.4}_{\pm 0.47}$	$\mathbf{7.03}_{\pm 0.67}$	$\mathbf{6.96}_{\pm 0.67}$
			Length/Maturity			
LLM sampling	$2.15_{\pm 0.73}$	$4.25_{\pm 1.01}$	$5.4_{\pm 2.1}$	$2.53_{\pm 1.12}$	$3.7_{\pm 1.4}$	$3.32_{\pm 1.12}$
GA	$2.66_{\pm 0.79}$	$2.81_{\pm 0.86}$	$3.17_{\pm 1.11}$	$2.27_{\pm 0.95}$	$1.94_{\pm 0.83}$	$2.08_{\pm 0.75}$
QD init only	$5.61_{\pm 1.46}$	$5.25_{\pm 1.14}$	$7.35_{\pm 0.6}$	$4.91_{\pm 1.39}$	$6.39_{\pm 1.06}$	$5.11_{\pm 1.49}$
QD-quality	$3.97_{\pm 1.36}$	$4.33_{\pm 1.18}$	$6.53_{\pm 1.13}$	$4.17_{\pm 1.17}$	$5.23_{\pm 1.61}$	$4.36_{\pm 1.59}$
QD-diversity	$\mathbf{7.32}_{\pm 0.82}$	$\mathbf{7.27}_{\pm 0.64}$	$\mathbf{7.67}_{\pm 0.27}$	$\mathbf{7.5}_{\pm 0.42}$	$\mathbf{7.28}_{\pm 0.27}$	$\mathbf{7.27}_{\pm 0.38}$
			Formality/Maturity			
LLM sampling	$1.8_{\pm 1.05}$	$4.39_{\pm 1.05}$	$4.74_{\pm 1.54}$	$2.86_{\pm 1.11}$	$3.61_{\pm 1.53}$	$3.11_{\pm 1.05}$
GA	$1.85_{\pm 0.83}$	$2.1_{\pm 0.9}$	$2.45_{\pm 1.02}$	$2.28_{\pm 0.96}$	$2.13_{\pm 0.94}$	$2.06_{\pm 0.71}$
QD init only	$4.88_{\pm 1.17}$	$5.32_{\pm 1.2}$	$\mathbf{5.58}_{\pm 1.16}$	$3.63_{\pm 1.27}$	$4.87_{\pm 1.07}$	$3.87_{\pm 1.15}$
QD-quality	$3.51_{\pm 0.89}$	$4.68_{\pm 1.33}$	$5.28_{\pm 1.39}$	$3.65_{\pm 1.21}$	$4.63_{\pm 1.42}$	$3.63_{\pm 1.37}$
QD-diversity	$\mathbf{6.31}_{\pm 1.05}$	$\mathbf{6.6}_{\pm 0.79}$	$4.69_{\pm 1.03}$	$\mathbf{5.73}_{\pm 1.22}$	$\mathbf{5.33}_{\pm 1.09}$	$\mathbf{4.62}_{\pm 0.97}$

Table I.11: The Max ROUGE-L results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.22_{\pm 0.06}$	$0.19_{\pm 0.07}$	$0.22_{\pm 0.09}$	$\mathbf{0.36}_{\pm 0.24}$	$0.14_{\pm 0.03}$	$0.17_{\pm 0.11}$
GA	$0.18_{\pm 0.06}$	$0.15_{\pm 0.06}$	$0.22_{\pm 0.09}$	$0.26_{\pm 0.21}$	$0.14_{\pm 0.02}$	$0.12_{\pm 0.07}$
QD init only	$\mathbf{0.23}_{\pm 0.06}$	$\mathbf{0.19}_{\pm 0.09}$	$\mathbf{0.31}_{\pm 0.1}$	$0.34_{\pm 0.22}$	$\mathbf{0.15}_{\pm 0.02}$	$\mathbf{0.17}_{\pm 0.1}$
QD-quality	$0.2_{\pm 0.05}$	$0.14_{\pm 0.06}$	$0.25_{\pm 0.08}$	$0.24_{\pm 0.18}$	$0.14_{\pm 0.02}$	$0.16_{\pm 0.1}$
QD-diversity	$0.22_{\pm 0.07}$	$0.17_{\pm 0.08}$	$0.31_{\pm 0.12}$	$0.3_{\pm 0.22}$	$0.14_{\pm 0.02}$	$0.14_{\pm 0.08}$
			Length/Maturity			
LLM sampling	$0.22_{\pm 0.06}$	$0.17_{\pm 0.06}$	$0.25_{\pm 0.12}$	$\mathbf{0.37}_{\pm 0.24}$	$0.14_{\pm 0.02}$	$0.15_{\pm 0.1}$
GA	$0.18_{\pm 0.06}$	$0.13_{\pm 0.06}$	$0.21_{\pm 0.1}$	$0.25_{\pm 0.19}$	$0.13_{\pm 0.02}$	$0.1_{\pm 0.06}$
QD init only	$\mathbf{0.23}_{\pm 0.07}$	$\mathbf{0.21}_{\pm 0.07}$	$\mathbf{0.32}_{\pm 0.1}$	$0.34_{\pm 0.23}$	$\mathbf{0.15}_{\pm 0.02}$	$0.16_{\pm 0.09}$
QD-quality	$0.19_{\pm 0.06}$	$0.14_{\pm 0.06}$	$0.26_{\pm 0.09}$	$0.27_{\pm 0.2}$	$0.14_{\pm 0.02}$	$0.14_{\pm 0.09}$
QD-diversity	$0.21_{\pm 0.06}$	$0.18_{\pm 0.06}$	$0.31_{\pm 0.09}$	$0.3_{\pm 0.21}$	$0.14_{\pm 0.02}$	$\mathbf{0.17}_{\pm 0.09}$
			Formality/Maturity			
LLM sampling	$\mathbf{0.2}_{\pm 0.06}$	$\mathbf{0.17}_{\pm 0.08}$	$0.16_{\pm 0.05}$	$\mathbf{0.38}_{\pm 0.25}$	$0.14_{\pm 0.02}$	$\mathbf{0.17}_{\pm 0.12}$
GA	$0.16_{\pm 0.06}$	$0.11_{\pm 0.05}$	$0.14_{\pm 0.05}$	$0.27_{\pm 0.21}$	$0.13_{\pm 0.02}$	$0.12_{\pm 0.08}$
QD init only	$0.19_{\pm 0.05}$	$0.17_{\pm 0.07}$	$\mathbf{0.18}_{\pm 0.04}$	$0.35_{\pm 0.23}$	$\mathbf{0.14}_{\pm 0.02}$	$0.16_{\pm 0.11}$
QD-quality	$0.17_{\pm 0.05}$	$0.16_{\pm 0.06}$	$0.16_{\pm 0.04}$	$0.22_{\pm 0.2}$	$0.14_{\pm 0.02}$	$0.14_{\pm 0.08}$
QD-diversity	$0.17_{\pm 0.05}$	$0.16_{\pm 0.06}$	$0.14_{\pm 0.04}$	$0.3_{\pm 0.23}$	$0.14_{\pm 0.02}$	$0.14_{\pm 0.07}$

Table I.12: The QD ROUGE-L results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.64_{\pm 0.34}$	$\mathbf{0.83}_{\pm 0.37}$	$0.93_{\pm 0.44}$	$0.85_{\pm 0.6}$	$0.6_{\pm 0.24}$	$0.52_{\pm 0.31}$
GA	$0.49_{\pm 0.25}$	$0.37_{\pm 0.18}$	$0.56_{\pm 0.26}$	$0.55_{\pm 0.43}$	$0.4_{\pm 0.18}$	$0.3_{\pm 0.2}$
QD init only	$1.09_{\pm 0.43}$	$0.78_{\pm 0.35}$	$1.36_{\pm 0.35}$	$\mathbf{1.17}_{\pm 0.76}$	$\mathbf{0.85}_{\pm 0.21}$	$\mathbf{0.68}_{\pm 0.39}$
QD-quality	$0.8_{\pm 0.36}$	$0.52_{\pm 0.26}$	$1.12_{\pm 0.36}$	$0.81_{\pm 0.62}$	$0.84_{\pm 0.23}$	$0.6_{\pm 0.32}$
QD-diversity	$\mathbf{1.21}_{\pm 0.32}$	$0.8_{\pm 0.36}$	$\mathbf{1.36}_{\pm 0.33}$	$1.05_{\pm 0.7}$	$0.85_{\pm 0.18}$	$0.56_{\pm 0.32}$
			Length/Maturity			
LLM sampling	$0.47_{\pm 0.2}$	$0.55_{\pm 0.28}$	$1.0_{\pm 0.56}$	$0.75_{\pm 0.52}$	$0.52_{\pm 0.21}$	$0.38_{\pm 0.26}$
GA	$0.41_{\pm 0.17}$	$0.33_{\pm 0.18}$	$0.56_{\pm 0.29}$	$0.46_{\pm 0.38}$	$0.29_{\pm 0.13}$	$0.19_{\pm 0.13}$
QD init only	$0.95_{\pm 0.34}$	$0.74_{\pm 0.3}$	$\mathbf{1.47}_{\pm 0.34}$	$1.0_{\pm 0.66}$	$0.83_{\pm 0.2}$	$0.55_{\pm 0.39}$
QD-quality	$0.62_{\pm 0.26}$	$0.46_{\pm 0.26}$	$1.2_{\pm 0.37}$	$0.66_{\pm 0.51}$	$0.7_{\pm 0.24}$	$0.43_{\pm 0.3}$
QD-diversity	$\mathbf{1.05}_{\pm 0.3}$	$\mathbf{0.84}_{\pm 0.34}$	$1.47_{\pm 0.31}$	$\mathbf{1.01}_{\pm 0.64}$	$\mathbf{0.89}_{\pm 0.13}$	$\mathbf{0.7}_{\pm 0.36}$
			Formality/Maturity			
LLM sampling	$0.39_{\pm 0.25}$	$0.56_{\pm 0.33}$	$0.66_{\pm 0.28}$	$0.9_{\pm 0.69}$	$0.52_{\pm 0.25}$	$0.39_{\pm 0.25}$
GA	$0.3_{\pm 0.18}$	$0.24_{\pm 0.15}$	$0.33_{\pm 0.17}$	$0.56_{\pm 0.52}$	$0.31_{\pm 0.14}$	$0.2_{\pm 0.13}$
QD init only	$0.84_{\pm 0.27}$	$0.68_{\pm 0.37}$	$\mathbf{0.78}_{\pm 0.24}$	$0.97_{\pm 0.77}$	$0.69_{\pm 0.18}$	$0.46_{\pm 0.32}$
QD-quality	$0.57_{\pm 0.23}$	$0.58_{\pm 0.26}$	$0.69_{\pm 0.27}$	$0.57_{\pm 0.56}$	$0.66_{\pm 0.23}$	$0.37_{\pm 0.27}$
QD-diversity	$\mathbf{0.93}_{\pm 0.27}$	$\mathbf{0.78}_{\pm 0.31}$	$0.59_{\pm 0.21}$	$\mathbf{0.98}_{\pm 0.74}$	$\mathbf{0.74}_{\pm 0.18}$	$\mathbf{0.48}_{\pm 0.29}$

Table I.13: The Max METEOR results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.32_{\pm 0.08}$	$0.24_{\pm 0.09}$	$0.33_{\pm 0.11}$	$\mathbf{0.33}_{\pm 0.21}$	$0.14_{\pm 0.03}$	$0.28_{\pm 0.14}$
GA	$0.25_{\pm 0.09}$	$0.2_{\pm 0.05}$	$0.31_{\pm 0.11}$	$0.28_{\pm 0.18}$	$0.15_{\pm 0.04}$	$0.22_{\pm 0.1}$
QD init only	$\mathbf{0.33}_{\pm 0.07}$	$\mathbf{0.26}_{\pm 0.11}$	$\mathbf{0.37}_{\pm 0.11}$	$0.32_{\pm 0.19}$	$0.16_{\pm 0.04}$	$\mathbf{0.28}_{\pm 0.13}$
QD-quality	$0.28_{\pm 0.08}$	$0.19_{\pm 0.05}$	$0.33_{\pm 0.1}$	$0.27_{\pm 0.18}$	$\mathbf{0.17}_{\pm 0.03}$	$0.25_{\pm 0.11}$
QD-diversity	$0.31_{\pm 0.08}$	$0.23_{\pm 0.09}$	$0.34_{\pm 0.11}$	$0.3_{\pm 0.18}$	$0.16_{\pm 0.04}$	$0.25_{\pm 0.11}$
			Length/Maturity			
LLM sampling	$\mathbf{0.32}_{\pm 0.07}$	$0.21_{\pm 0.09}$	$0.34_{\pm 0.12}$	$\mathbf{0.34}_{\pm 0.21}$	$0.14_{\pm 0.03}$	$0.24_{\pm 0.13}$
GA	$0.24_{\pm 0.09}$	$0.19_{\pm 0.07}$	$0.3_{\pm 0.11}$	$0.28_{\pm 0.19}$	$0.14_{\pm 0.03}$	$0.19_{\pm 0.1}$
QD init only	$0.31_{\pm 0.08}$	$\mathbf{0.3}_{\pm 0.1}$	$\mathbf{0.37}_{\pm 0.12}$	$0.32_{\pm 0.19}$	$\mathbf{0.17}_{\pm 0.04}$	$0.25_{\pm 0.1}$
QD-quality	$0.26_{\pm 0.08}$	$0.2_{\pm 0.08}$	$0.35_{\pm 0.1}$	$0.28_{\pm 0.18}$	$0.16_{\pm 0.03}$	$0.23_{\pm 0.11}$
QD-diversity	$0.29_{\pm 0.09}$	$0.26_{\pm 0.09}$	$0.36_{\pm 0.1}$	$0.29_{\pm 0.18}$	$0.16_{\pm 0.03}$	$\mathbf{0.26}_{\pm 0.11}$
			Formality/Maturity			
LLM sampling	$0.31_{\pm 0.08}$	$0.19_{\pm 0.1}$	$0.3_{\pm 0.09}$	$\mathbf{0.34}_{\pm 0.22}$	$0.14_{\pm 0.04}$	$\mathbf{0.26}_{\pm 0.14}$
GA	$0.25_{\pm 0.09}$	$0.17_{\pm 0.05}$	$0.25_{\pm 0.08}$	$0.29_{\pm 0.19}$	$0.14_{\pm 0.04}$	$0.2_{\pm 0.1}$
QD init only	$\mathbf{0.33}_{\pm 0.08}$	$\mathbf{0.24}_{\pm 0.09}$	$\mathbf{0.31}_{\pm 0.07}$	$0.34_{\pm 0.21}$	$0.15_{\pm 0.04}$	$0.24_{\pm 0.12}$
QD-quality	$0.29_{\pm 0.07}$	$0.24_{\pm 0.08}$	$0.29_{\pm 0.08}$	$0.25_{\pm 0.17}$	$0.15_{\pm 0.04}$	$0.24_{\pm 0.12}$
QD-diversity	$0.27_{\pm 0.08}$	$0.23_{\pm 0.07}$	$0.28_{\pm 0.07}$	$0.31_{\pm 0.2}$	$\mathbf{0.16}_{\pm 0.04}$	$0.23_{\pm 0.11}$

Table I.14: The QD METEOR results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.89_{\pm 0.43}$	$1.04_{\pm 0.3}$	$1.5_{\pm 0.63}$	$0.88_{\pm 0.68}$	$0.53_{\pm 0.22}$	$0.92_{\pm 0.5}$
GA	$0.61_{\pm 0.36}$	$0.51_{\pm 0.17}$	$0.87_{\pm 0.34}$	$0.68_{\pm 0.52}$	$0.41_{\pm 0.19}$	$0.58_{\pm 0.3}$
QD init only	$1.41_{\pm 0.58}$	$1.07_{\pm 0.33}$	$\mathbf{1.97}_{\pm 0.48}$	$\mathbf{1.41}_{\pm 0.97}$	$0.79_{\pm 0.24}$	$\mathbf{1.25}_{\pm 0.54}$
QD-quality	$1.02_{\pm 0.47}$	$0.79_{\pm 0.22}$	$1.73_{\pm 0.52}$	$1.11_{\pm 0.82}$	$0.78_{\pm 0.24}$	$1.15_{\pm 0.51}$
QD-diversity	$\mathbf{1.55}_{\pm 0.46}$	$\mathbf{1.11}_{\pm 0.26}$	$1.9_{\pm 0.45}$	$1.36_{\pm 0.91}$	$\mathbf{0.81}_{\pm 0.22}$	$1.14_{\pm 0.45}$
			Length/Maturity			
LLM sampling	$0.67_{\pm 0.25}$	$0.67_{\pm 0.32}$	$1.56_{\pm 0.78}$	$0.78_{\pm 0.59}$	$0.45_{\pm 0.18}$	$0.65_{\pm 0.37}$
GA	$0.51_{\pm 0.22}$	$0.49_{\pm 0.21}$	$0.88_{\pm 0.4}$	$0.58_{\pm 0.46}$	$0.28_{\pm 0.12}$	$0.38_{\pm 0.23}$
QD init only	$1.2_{\pm 0.43}$	$1.02_{\pm 0.35}$	$2.09_{\pm 0.51}$	$1.17_{\pm 0.8}$	$0.78_{\pm 0.21}$	$0.98_{\pm 0.51}$
QD-quality	$0.77_{\pm 0.32}$	$0.7_{\pm 0.29}$	$1.84_{\pm 0.52}$	$0.8_{\pm 0.59}$	$0.65_{\pm 0.23}$	$0.78_{\pm 0.45}$
QD-diversity	$\mathbf{1.34}_{\pm 0.42}$	$\mathbf{1.17}_{\pm 0.3}$	$\mathbf{2.09}_{\pm 0.47}$	$\mathbf{1.28}_{\pm 0.82}$	$\mathbf{0.85}_{\pm 0.19}$	$\mathbf{1.28}_{\pm 0.5}$
			Formality/Maturity			
LLM sampling	$0.57_{\pm 0.36}$	$0.67_{\pm 0.36}$	$1.27_{\pm 0.52}$	$0.89_{\pm 0.68}$	$0.48_{\pm 0.24}$	$0.65_{\pm 0.35}$
GA	$0.47_{\pm 0.26}$	$0.36_{\pm 0.18}$	$0.62_{\pm 0.31}$	$0.62_{\pm 0.52}$	$0.31_{\pm 0.15}$	$0.38_{\pm 0.2}$
QD init only	$1.43_{\pm 0.43}$	$0.97_{\pm 0.39}$	$\mathbf{1.47}_{\pm 0.41}$	$1.02_{\pm 0.77}$	$0.65_{\pm 0.2}$	$0.74_{\pm 0.42}$
QD-quality	$0.94_{\pm 0.32}$	$0.89_{\pm 0.32}$	$1.34_{\pm 0.44}$	$0.71_{\pm 0.58}$	$0.65_{\pm 0.25}$	$0.7_{\pm 0.4}$
QD-diversity	$\mathbf{1.49}_{\pm 0.45}$	$\mathbf{1.2}_{\pm 0.35}$	$1.16_{\pm 0.36}$	$\mathbf{1.2}_{\pm 0.89}$	$\mathbf{0.77}_{\pm 0.22}$	$\mathbf{0.86}_{\pm 0.46}$

Table I.15: The Max BLEU results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.07_{\pm 0.04}$	$0.08_{\pm 0.02}$	$0.08_{\pm 0.05}$	$\mathbf{0.27}_{\pm 0.2}$	$0.02_{\pm 0.01}$	$0.08_{\pm 0.04}$
GA	$0.06_{\pm 0.03}$	$0.06_{\pm 0.02}$	$0.07_{\pm 0.04}$	$0.14_{\pm 0.13}$	$0.02_{\pm 0.01}$	$0.06_{\pm 0.04}$
QD init only	$\mathbf{0.08}_{\pm 0.05}$	$\mathbf{0.09}_{\pm 0.03}$	$0.12_{\pm 0.04}$	$0.25_{\pm 0.18}$	$\mathbf{0.02}_{\pm 0.01}$	$0.09_{\pm 0.05}$
QD-quality	$0.06_{\pm 0.04}$	$0.07_{\pm 0.02}$	$0.09_{\pm 0.03}$	$0.14_{\pm 0.1}$	$0.02_{\pm 0.01}$	$0.08_{\pm 0.04}$
QD-diversity	$0.07_{\pm 0.03}$	$0.09_{\pm 0.02}$	$\mathbf{0.12}_{\pm 0.05}$	$0.21_{\pm 0.16}$	$0.02_{\pm 0.01}$	$\mathbf{0.11}_{\pm 0.06}$
			Length/Maturity			
LLM sampling	$0.07_{\pm 0.04}$	$0.08_{\pm 0.03}$	$0.09_{\pm 0.06}$	$\mathbf{0.27}_{\pm 0.19}$	$0.02_{\pm 0.01}$	$0.08_{\pm 0.05}$
GA	$0.05_{\pm 0.03}$	$0.06_{\pm 0.03}$	$0.07_{\pm 0.04}$	$0.13_{\pm 0.12}$	$0.02_{\pm 0.01}$	$0.05_{\pm 0.02}$
QD init only	$\mathbf{0.08}_{\pm 0.04}$	$0.09_{\pm 0.03}$	$\mathbf{0.13}_{\pm 0.05}$	$0.22_{\pm 0.18}$	$\mathbf{0.03}_{\pm 0.01}$	$\mathbf{0.11}_{\pm 0.08}$
QD-quality	$0.06_{\pm 0.03}$	$0.07_{\pm 0.02}$	$0.1_{\pm 0.06}$	$0.16_{\pm 0.14}$	$0.02_{\pm 0.01}$	$0.09_{\pm 0.05}$
QD-diversity	$0.07_{\pm 0.03}$	$\mathbf{0.09}_{\pm 0.03}$	$0.12_{\pm 0.05}$	$0.2_{\pm 0.13}$	$0.02_{\pm 0.01}$	$0.09_{\pm 0.05}$
			Formality/Maturity			
LLM sampling	$\mathbf{0.06}_{\pm 0.04}$	$0.08_{\pm 0.02}$	$0.04_{\pm 0.02}$	$\mathbf{0.3}_{\pm 0.2}$	$0.02_{\pm 0.01}$	$0.08_{\pm 0.06}$
GA	$0.04_{\pm 0.03}$	$0.05_{\pm 0.02}$	$0.04_{\pm 0.02}$	$0.17_{\pm 0.17}$	$0.02_{\pm 0.01}$	$0.05_{\pm 0.04}$
QD init only	$0.06_{\pm 0.03}$	$\mathbf{0.08}_{\pm 0.03}$	$\mathbf{0.05}_{\pm 0.02}$	$0.26_{\pm 0.21}$	$0.02_{\pm 0.01}$	$\mathbf{0.08}_{\pm 0.04}$
QD-quality	$0.05_{\pm 0.03}$	$0.07_{\pm 0.03}$	$0.04_{\pm 0.01}$	$0.14_{\pm 0.15}$	$0.02_{\pm 0.01}$	$0.06_{\pm 0.04}$
QD-diversity	$0.04_{\pm 0.02}$	$0.07_{\pm 0.02}$	$0.04_{\pm 0.01}$	$0.2_{\pm 0.17}$	$\mathbf{0.02}_{\pm 0.01}$	$0.07_{\pm 0.04}$

Table I.16: The QD BLEU results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.18_{\pm 0.12}$	$0.34_{\pm 0.06}$	$0.29_{\pm 0.17}$	$0.5_{\pm 0.31}$	$0.06_{\pm 0.04}$	$0.25_{\pm 0.11}$
GA	$0.12_{\pm 0.09}$	$0.15_{\pm 0.07}$	$0.17_{\pm 0.09}$	$0.26_{\pm 0.19}$	$0.05_{\pm 0.03}$	$0.15_{\pm 0.07}$
QD init only	$0.29_{\pm 0.16}$	$0.33_{\pm 0.07}$	$\mathbf{0.45}_{\pm 0.11}$	$\mathbf{0.58}_{\pm 0.3}$	$\mathbf{0.09}_{\pm 0.03}$	$0.35_{\pm 0.11}$
QD-quality	$0.2_{\pm 0.14}$	$0.24_{\pm 0.08}$	$0.35_{\pm 0.11}$	$0.37_{\pm 0.2}$	$0.09_{\pm 0.04}$	$0.31_{\pm 0.09}$
QD-diversity	$\mathbf{0.32}_{\pm 0.13}$	$\mathbf{0.37}_{\pm 0.06}$	$0.45_{\pm 0.11}$	$0.57_{\pm 0.24}$	$0.09_{\pm 0.03}$	$\mathbf{0.37}_{\pm 0.09}$
			Length/Maturity			
LLM sampling	$0.14_{\pm 0.08}$	$0.25_{\pm 0.07}$	$0.31_{\pm 0.2}$	$0.45_{\pm 0.28}$	$0.05_{\pm 0.03}$	$0.2_{\pm 0.09}$
GA	$0.1_{\pm 0.05}$	$0.12_{\pm 0.05}$	$0.17_{\pm 0.1}$	$0.21_{\pm 0.17}$	$0.03_{\pm 0.02}$	$0.1_{\pm 0.05}$
QD init only	$0.25_{\pm 0.12}$	$0.29_{\pm 0.07}$	$\mathbf{0.49}_{\pm 0.12}$	$0.5_{\pm 0.3}$	$0.09_{\pm 0.03}$	$0.31_{\pm 0.13}$
QD-quality	$0.16_{\pm 0.08}$	$0.19_{\pm 0.08}$	$0.38_{\pm 0.13}$	$0.32_{\pm 0.21}$	$0.07_{\pm 0.04}$	$0.24_{\pm 0.1}$
QD-diversity	$\mathbf{0.27}_{\pm 0.1}$	$\mathbf{0.38}_{\pm 0.09}$	$0.47_{\pm 0.09}$	$\mathbf{0.57}_{\pm 0.23}$	$\mathbf{0.09}_{\pm 0.03}$	$\mathbf{0.37}_{\pm 0.09}$
			Formality/Maturity			
LLM sampling	$0.11_{\pm 0.09}$	$0.25_{\pm 0.1}$	$0.17_{\pm 0.07}$	$\mathbf{0.6}_{\pm 0.37}$	$0.06_{\pm 0.04}$	$0.21_{\pm 0.13}$
GA	$0.08_{\pm 0.05}$	$0.09_{\pm 0.05}$	$0.08_{\pm 0.04}$	$0.31_{\pm 0.31}$	$0.04_{\pm 0.03}$	$0.11_{\pm 0.07}$
QD init only	$0.2_{\pm 0.09}$	$0.27_{\pm 0.11}$	$\mathbf{0.2}_{\pm 0.06}$	$0.59_{\pm 0.43}$	$0.08_{\pm 0.05}$	$\mathbf{0.24}_{\pm 0.12}$
QD-quality	$0.14_{\pm 0.08}$	$0.2_{\pm 0.08}$	$0.17_{\pm 0.07}$	$0.33_{\pm 0.35}$	$0.08_{\pm 0.05}$	$0.2_{\pm 0.11}$
QD-diversity	$\mathbf{0.21}_{\pm 0.07}$	$\mathbf{0.28}_{\pm 0.09}$	$0.15_{\pm 0.05}$	$0.5_{\pm 0.37}$	$\mathbf{0.09}_{\pm 0.05}$	$0.24_{\pm 0.1}$

Table I.17: The Max CHRF results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$38.17_{\pm 5.16}$	$24.41_{\pm 6.55}$	$30.68_{\pm 6.39}$	$\mathbf{41.03}_{\pm 13.89}$	$25.0_{\pm 3.99}$	$\mathbf{21.13}_{\pm 10.17}$
GA	$33.05_{\pm 6.14}$	$21.94_{\pm 5.59}$	$30.72_{\pm 7.17}$	$31.77_{\pm 13.65}$	$28.15_{\pm 3.85}$	$17.17_{\pm 7.6}$
QD init only	$\mathbf{38.44}_{\pm 5.23}$	$\mathbf{26.94}_{\pm 8.28}$	$\mathbf{35.39}_{\pm 7.03}$	$38.87_{\pm 12.86}$	$28.8_{\pm 3.69}$	$20.97_{\pm 9.35}$
QD-quality	$34.89_{\pm 5.35}$	$22.51_{\pm 6.66}$	$33.19_{\pm 6.18}$	$32.11_{\pm 13.49}$	$\mathbf{29.28}_{\pm 3.33}$	$19.06_{\pm 8.23}$
QD-diversity	$36.7_{\pm 4.84}$	$25.38_{\pm 8.12}$	$34.48_{\pm 7.37}$	$35.4_{\pm 12.73}$	$28.97_{\pm 4.09}$	$18.74_{\pm 7.49}$
			Length/Maturity			
LLM sampling	$\mathbf{37.84}_{\pm 5.22}$	$22.1_{\pm 6.87}$	$32.58_{\pm 8.01}$	$\mathbf{41.33}_{\pm 13.69}$	$25.06_{\pm 4.04}$	$\mathbf{20.62}_{\pm 10.52}$
GA	$32.0_{\pm 5.39}$	$21.33_{\pm 5.06}$	$30.7_{\pm 7.48}$	$32.43_{\pm 15.25}$	$26.42_{\pm 3.66}$	$16.26_{\pm 6.82}$
QD init only	$37.05_{\pm 5.26}$	$\mathbf{29.7}_{\pm 7.66}$	$\mathbf{36.23}_{\pm 7.35}$	$38.3_{\pm 12.43}$	$28.2_{\pm 3.88}$	$20.45_{\pm 8.55}$
QD-quality	$33.6_{\pm 5.54}$	$22.78_{\pm 5.61}$	$33.95_{\pm 6.56}$	$33.04_{\pm 13.92}$	$\mathbf{28.54}_{\pm 3.42}$	$18.17_{\pm 8.22}$
QD-diversity	$35.29_{\pm 6.13}$	$26.22_{\pm 7.57}$	$34.58_{\pm 6.29}$	$34.7_{\pm 12.21}$	$27.81_{\pm 3.47}$	$20.37_{\pm 8.76}$
			Formality/Maturity			
LLM sampling	$\mathbf{36.93}_{\pm 5.64}$	$23.38_{\pm 7.32}$	$27.49_{\pm 6.14}$	$\mathbf{41.8}_{\pm 13.77}$	$24.26_{\pm 4.27}$	$\mathbf{21.07}_{\pm 10.98}$
GA	$31.89_{\pm 5.59}$	$19.2_{\pm 4.25}$	$24.95_{\pm 5.73}$	$34.1_{\pm 15.44}$	$25.14_{\pm 4.63}$	$16.34_{\pm 7.1}$
QD init only	$36.5_{\pm 4.82}$	$24.74_{\pm 7.96}$	$\mathbf{28.76}_{\pm 5.51}$	$40.68_{\pm 13.92}$	$25.26_{\pm 4.06}$	$20.17_{\pm 9.15}$
QD-quality	$34.07_{\pm 4.81}$	$\mathbf{24.98}_{\pm 7.24}$	$26.63_{\pm 5.44}$	$29.55_{\pm 14.89}$	$26.02_{\pm 4.31}$	$18.87_{\pm 8.9}$
QD-diversity	$32.65_{\pm 5.02}$	$24.78_{\pm 7.85}$	$26.07_{\pm 4.89}$	$37.55_{\pm 14.14}$	$\mathbf{26.77}_{\pm 4.43}$	$19.04_{\pm 7.28}$

Table I.18: The QD CHRF results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$117.47_{\pm 53.98}$	$136.92_{\pm 29.53}$	$158.95_{\pm 62.18}$	$109.18_{\pm 56.17}$	$102.31_{\pm 39.72}$	$84.04_{\pm 46.7}$
GA	$90.91_{\pm 42.53}$	$66.16_{\pm 26.51}$	$94.5_{\pm 35.85}$	$79.94_{\pm 42.04}$	$77.48_{\pm 32.41}$	$52.26_{\pm 28.83}$
QD init only	$187.45_{\pm 67.43}$	$140.51_{\pm 34.88}$	$\mathbf{215.59}_{\pm 48.62}$	$174.09_{\pm 81.36}$	$151.44_{\pm 34.24}$	$\mathbf{114.39}_{\pm 51.45}$
QD-quality	$151.14_{\pm 59.15}$	$104.0_{\pm 24.61}$	$189.23_{\pm 49.8}$	$134.15_{\pm 70.04}$	$151.61_{\pm 38.31}$	$104.97_{\pm 47.02}$
QD-diversity	$\mathbf{224.72}_{\pm 45.18}$	$\mathbf{156.77}_{\pm 32.59}$	$215.44_{\pm 46.49}$	$\mathbf{174.76}_{\pm 72.15}$	$\mathbf{160.59}_{\pm 26.44}$	$111.87_{\pm 47.38}$
			Length/Maturity			
LLM sampling	$86.4_{\pm 29.37}$	$89.77_{\pm 33.9}$	$164.06_{\pm 76.34}$	$94.76_{\pm 52.59}$	$86.98_{\pm 33.84}$	$64.42_{\pm 35.84}$
GA	$77.73_{\pm 24.05}$	$60.91_{\pm 21.3}$	$95.13_{\pm 40.23}$	$67.8_{\pm 40.63}$	$54.39_{\pm 22.41}$	$35.51_{\pm 18.94}$
QD init only	$162.19_{\pm 50.68}$	$125.8_{\pm 39.53}$	$\mathbf{232.78}_{\pm 44.83}$	$144.3_{\pm 67.99}$	$139.4_{\pm 31.01}$	$93.62_{\pm 49.19}$
QD-quality	$116.22_{\pm 42.1}$	$88.2_{\pm 32.12}$	$199.78_{\pm 50.51}$	$101.93_{\pm 51.35}$	$123.23_{\pm 40.12}$	$75.12_{\pm 41.37}$
QD-diversity	$\mathbf{197.18}_{\pm 42.07}$	$\mathbf{153.21}_{\pm 38.08}$	$231.65_{\pm 40.82}$	$\mathbf{171.27}_{\pm 64.4}$	$\mathbf{160.99}_{\pm 21.79}$	$\mathbf{122.58}_{\pm 46.28}$
			Formality/Maturity			
LLM sampling	$73.48_{\pm 44.46}$	$92.47_{\pm 33.6}$	$127.44_{\pm 50.31}$	$110.13_{\pm 56.9}$	$89.8_{\pm 42.34}$	$61.64_{\pm 30.92}$
GA	$64.39_{\pm 30.45}$	$43.45_{\pm 19.26}$	$63.82_{\pm 29.18}$	$76.47_{\pm 49.9}$	$59.37_{\pm 27.71}$	$35.03_{\pm 16.61}$
QD init only	$177.18_{\pm 47.52}$	$119.2_{\pm 42.55}$	$\mathbf{152.63}_{\pm 44.34}$	$131.21_{\pm 65.69}$	$118.05_{\pm 34.94}$	$74.48_{\pm 40.93}$
QD-quality	$125.16_{\pm 38.03}$	$102.54_{\pm 32.24}$	$134.36_{\pm 45.31}$	$92.78_{\pm 57.48}$	$121.55_{\pm 44.47}$	$66.7_{\pm 39.74}$
QD-diversity	$\mathbf{204.36}_{\pm 42.23}$	$\mathbf{144.17}_{\pm 33.02}$	$120.3_{\pm 33.58}$	$\mathbf{153.54}_{\pm 68.05}$	$\mathbf{138.6}_{\pm 38.74}$	$\mathbf{82.76}_{\pm 40.7}$

Table I.19: The Max SacreBLEU results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$5.46_{\pm 3.5}$	$3.56_{\pm 1.44}$	$3.68_{\pm 2.5}$	$\mathbf{12.75}_{\pm 13.87}$	$1.2_{\pm 0.72}$	$3.44_{\pm 3.0}$
GA	$3.84_{\pm 2.84}$	$2.62_{\pm 1.48}$	$3.35_{\pm 3.13}$	$6.66_{\pm 8.45}$	$1.27_{\pm 0.87}$	$2.22_{\pm 1.82}$
QD init only	$\mathbf{6.43}_{\pm 4.67}$	$\mathbf{4.56}_{\pm 3.63}$	$\mathbf{5.88}_{\pm 3.27}$	$9.53_{\pm 9.19}$	$\mathbf{1.6}_{\pm 0.82}$	$\mathbf{3.98}_{\pm 2.91}$
QD-quality	$4.56_{\pm 3.62}$	$2.79_{\pm 1.1}$	$4.2_{\pm 2.63}$	$6.13_{\pm 6.84}$	$1.49_{\pm 0.86}$	$2.81_{\pm 1.91}$
QD-diversity	$5.19_{\pm 3.2}$	$4.02_{\pm 2.03}$	$5.77_{\pm 3.98}$	$8.38_{\pm 9.74}$	$1.52_{\pm 0.8}$	$3.76_{\pm 2.19}$
			Length/Maturity			
LLM sampling	$5.49_{\pm 3.61}$	$3.59_{\pm 1.32}$	$4.43_{\pm 3.87}$	$\mathbf{12.7}_{\pm 13.96}$	$1.21_{\pm 0.65}$	$3.38_{\pm 3.73}$
GA	$3.18_{\pm 2.42}$	$2.13_{\pm 1.08}$	$3.2_{\pm 2.39}$	$7.93_{\pm 11.1}$	$1.06_{\pm 0.47}$	$1.79_{\pm 0.97}$
QD init only	$\mathbf{5.92}_{\pm 3.95}$	$\mathbf{4.51}_{\pm 2.67}$	$\mathbf{6.41}_{\pm 4.95}$	$9.9_{\pm 10.92}$	$\mathbf{1.68}_{\pm 0.78}$	$\mathbf{4.07}_{\pm 3.09}$
QD-quality	$4.08_{\pm 3.0}$	$2.59_{\pm 0.96}$	$5.14_{\pm 4.66}$	$7.24_{\pm 9.6}$	$1.63_{\pm 1.33}$	$3.46_{\pm 2.67}$
QD-diversity	$4.6_{\pm 2.99}$	$4.4_{\pm 2.63}$	$5.64_{\pm 3.77}$	$9.17_{\pm 10.03}$	$1.43_{\pm 0.58}$	$3.83_{\pm 2.9}$
			Formality/Maturity			
LLM sampling	$\mathbf{4.53}_{\pm 3.61}$	$3.36_{\pm 1.6}$	$2.12_{\pm 1.2}$	$\mathbf{14.59}_{\pm 14.77}$	$1.12_{\pm 0.59}$	$\mathbf{3.62}_{\pm 4.7}$
GA	$3.08_{\pm 2.66}$	$1.89_{\pm 1.07}$	$1.61_{\pm 1.01}$	$8.79_{\pm 11.94}$	$1.13_{\pm 1.14}$	$1.97_{\pm 1.94}$
QD init only	$4.33_{\pm 2.57}$	$\mathbf{3.8}_{\pm 1.34}$	$\mathbf{2.35}_{\pm 1.27}$	$12.02_{\pm 14.11}$	$1.32_{\pm 1.07}$	$3.4_{\pm 2.77}$
QD-quality	$3.46_{\pm 2.73}$	$3.22_{\pm 3.0}$	$1.83_{\pm 0.99}$	$6.24_{\pm 9.58}$	$1.16_{\pm 0.69}$	$2.59_{\pm 2.12}$
QD-diversity	$2.97_{\pm 2.05}$	$2.65_{\pm 0.9}$	$1.78_{\pm 0.79}$	$9.9_{\pm 11.5}$	$\mathbf{1.34}_{\pm 0.74}$	$2.65_{\pm 2.03}$

Table I.20: The QD SacreBLEU results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$14.22_{\pm 9.91}$	$11.31_{\pm 2.11}$	$6.38_{\pm 3.5}$	$23.26_{\pm 22.33}$	$3.28_{\pm 2.17}$	$9.98_{\pm 6.73}$
GA	$9.04_{\pm 5.18}$	$6.4_{\pm 3.49}$	$6.83_{\pm 3.28}$	$13.17_{\pm 11.54}$	$2.55_{\pm 1.62}$	$5.76_{\pm 3.75}$
QD init only	$\mathbf{22.5}_{\pm 14.05}$	$12.63_{\pm 3.69}$	$\mathbf{19.22}_{\pm 7.01}$	$\mathbf{26.83}_{\pm 22.87}$	$\mathbf{5.22}_{\pm 2.4}$	$\mathbf{13.51}_{\pm 7.13}$
QD-quality	$12.47_{\pm 9.02}$	$7.17_{\pm 2.06}$	$14.67_{\pm 7.61}$	$12.08_{\pm 11.31}$	$4.77_{\pm 2.1}$	$10.76_{\pm 5.81}$
QD-diversity	$21.0_{\pm 11.69}$	$\mathbf{13.93}_{\pm 3.1}$	$18.31_{\pm 7.17}$	$23.23_{\pm 18.58}$	$4.83_{\pm 1.23}$	$13.24_{\pm 5.41}$
			Length/Maturity			
LLM sampling	$10.07_{\pm 6.74}$	$8.59_{\pm 2.36}$	$5.81_{\pm 3.27}$	$18.13_{\pm 18.1}$	$2.8_{\pm 1.68}$	$8.86_{\pm 8.25}$
GA	$6.81_{\pm 4.12}$	$4.05_{\pm 2.33}$	$7.11_{\pm 4.12}$	$11.92_{\pm 13.21}$	$1.7_{\pm 0.98}$	$3.93_{\pm 2.31}$
QD init only	$\mathbf{16.82}_{\pm 10.89}$	$12.92_{\pm 4.69}$	$\mathbf{21.64}_{\pm 8.24}$	$\mathbf{24.15}_{\pm 23.03}$	$\mathbf{5.26}_{\pm 2.1}$	$12.79_{\pm 9.77}$
QD-quality	$9.36_{\pm 6.69}$	$6.33_{\pm 2.66}$	$16.97_{\pm 8.33}$	$10.48_{\pm 9.72}$	$4.26_{\pm 2.68}$	$9.29_{\pm 5.99}$
QD-diversity	$16.43_{\pm 8.71}$	$\mathbf{14.96}_{\pm 5.04}$	$18.54_{\pm 5.44}$	$23.93_{\pm 20.34}$	$5.18_{\pm 1.64}$	$\mathbf{14.64}_{\pm 7.82}$
			Formality/Maturity			
LLM sampling	$7.81_{\pm 6.18}$	$7.75_{\pm 2.02}$	$5.72_{\pm 3.8}$	$27.16_{\pm 23.93}$	$3.07_{\pm 2.13}$	$9.12_{\pm 8.39}$
GA	$5.55_{\pm 4.74}$	$3.32_{\pm 2.0}$	$3.39_{\pm 1.86}$	$13.0_{\pm 13.47}$	$2.02_{\pm 1.58}$	$3.72_{\pm 3.02}$
QD init only	$\mathbf{12.81}_{\pm 7.11}$	$8.9_{\pm 3.04}$	$\mathbf{9.11}_{\pm 4.07}$	$\mathbf{28.51}_{\pm 27.62}$	$4.3_{\pm 3.08}$	$\mathbf{10.77}_{\pm 8.92}$
QD-quality	$8.57_{\pm 6.92}$	$\mathbf{9.29}_{\pm 7.94}$	$6.69_{\pm 2.93}$	$7.31_{\pm 8.55}$	$4.07_{\pm 3.25}$	$8.23_{\pm 6.65}$
QD-diversity	$11.35_{\pm 5.46}$	$8.79_{\pm 2.37}$	$5.79_{\pm 2.52}$	$24.04_{\pm 23.01}$	$\mathbf{4.9}_{\pm 3.04}$	$9.62_{\pm 8.46}$

Table I.21: The Self-BLEU results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.59_{\pm 0.07}$	$0.44_{\pm 0.04}$	$0.42_{\pm 0.09}$	$0.76_{\pm 0.23}$	$0.33_{\pm 0.09}$	$0.42_{\pm 0.13}$
GA	$0.85_{\pm 0.14}$	$0.8_{\pm 0.15}$	$0.85_{\pm 0.1}$	$0.88_{\pm 0.14}$	$0.89_{\pm 0.13}$	$0.87_{\pm 0.12}$
QD init only	$0.46_{\pm 0.07}$	$0.26_{\pm 0.05}$	$0.22_{\pm 0.08}$	$0.55_{\pm 0.1}$	$0.26_{\pm 0.08}$	$0.26_{\pm 0.1}$
QD-quality	$0.56_{\pm 0.28}$	$0.53_{\pm 0.1}$	$0.42_{\pm 0.21}$	$0.65_{\pm 0.17}$	$0.37_{\pm 0.27}$	$0.4_{\pm 0.23}$
QD-diversity	$\mathbf{0.2}_{\pm 0.09}$	$\mathbf{0.12}_{\pm 0.05}$	$\mathbf{0.19}_{\pm 0.16}$	$\mathbf{0.21}_{\pm 0.12}$	$\mathbf{0.15}_{\pm 0.13}$	$\mathbf{0.21}_{\pm 0.15}$
			Length/Maturity			
LLM sampling	$0.59_{\pm 0.07}$	$0.43_{\pm 0.07}$	$0.41_{\pm 0.1}$	$0.75_{\pm 0.24}$	$0.35_{\pm 0.09}$	$0.4_{\pm 0.13}$
GA	$0.91_{\pm 0.1}$	$0.9_{\pm 0.09}$	$0.86_{\pm 0.13}$	$0.91_{\pm 0.1}$	$0.94_{\pm 0.08}$	$0.93_{\pm 0.11}$
QD init only	$0.46_{\pm 0.07}$	$0.27_{\pm 0.06}$	$0.25_{\pm 0.08}$	$0.61_{\pm 0.11}$	$0.25_{\pm 0.09}$	$0.29_{\pm 0.13}$
QD-quality	$0.8_{\pm 0.19}$	$0.7_{\pm 0.14}$	$0.38_{\pm 0.19}$	$0.77_{\pm 0.14}$	$0.57_{\pm 0.28}$	$0.64_{\pm 0.26}$
QD-diversity	$\mathbf{0.27}_{\pm 0.19}$	$\mathbf{0.22}_{\pm 0.16}$	$\mathbf{0.1}_{\pm 0.09}$	$\mathbf{0.2}_{\pm 0.12}$	$\mathbf{0.14}_{\pm 0.11}$	$\mathbf{0.17}_{\pm 0.1}$
			Formality/Maturity			
LLM sampling	$0.59_{\pm 0.07}$	$0.43_{\pm 0.05}$	$0.41_{\pm 0.09}$	$0.76_{\pm 0.22}$	$0.34_{\pm 0.1}$	$0.4_{\pm 0.13}$
GA	$0.94_{\pm 0.06}$	$0.93_{\pm 0.1}$	$0.91_{\pm 0.11}$	$0.93_{\pm 0.08}$	$0.95_{\pm 0.08}$	$0.94_{\pm 0.07}$
QD init only	$\mathbf{0.47}_{\pm 0.04}$	$\mathbf{0.34}_{\pm 0.04}$	$\mathbf{0.26}_{\pm 0.07}$	$0.7_{\pm 0.15}$	$\mathbf{0.26}_{\pm 0.07}$	$\mathbf{0.33}_{\pm 0.11}$
QD-quality	$0.84_{\pm 0.11}$	$0.68_{\pm 0.22}$	$0.6_{\pm 0.2}$	$0.81_{\pm 0.11}$	$0.65_{\pm 0.25}$	$0.77_{\pm 0.21}$
QD-diversity	$0.54_{\pm 0.15}$	$0.36_{\pm 0.2}$	$0.66_{\pm 0.17}$	$\mathbf{0.53}_{\pm 0.2}$	$0.56_{\pm 0.2}$	$0.66_{\pm 0.15}$

Table I.22: The Distinct-1 results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.12_{\pm 0.02}$	$0.14_{\pm 0.02}$	$0.19_{\pm 0.03}$	$0.1_{\pm 0.04}$	$0.23_{\pm 0.03}$	$0.19_{\pm 0.03}$
GA	$0.23_{\pm 0.07}$	$0.19_{\pm 0.06}$	$0.24_{\pm 0.07}$	$0.21_{\pm 0.07}$	$0.2_{\pm 0.07}$	$0.24_{\pm 0.07}$
QD init only	$0.17_{\pm 0.02}$	$0.21_{\pm 0.02}$	$0.27_{\pm 0.03}$	$0.18_{\pm 0.02}$	$0.25_{\pm 0.06}$	$0.24_{\pm 0.03}$
QD-quality	$0.32_{\pm 0.09}$	$0.29_{\pm 0.08}$	$0.39_{\pm 0.07}$	$0.3_{\pm 0.07}$	$0.38_{\pm 0.1}$	$0.37_{\pm 0.08}$
QD-diversity	$\mathbf{0.41}_{\pm 0.05}$	$\mathbf{0.42}_{\pm 0.04}$	$\mathbf{0.46}_{\pm 0.05}$	$\mathbf{0.43}_{\pm 0.05}$	$\mathbf{0.45}_{\pm 0.09}$	$\mathbf{0.45}_{\pm 0.06}$
			Length/Maturity			
LLM sampling	$0.13_{\pm 0.02}$	$0.14_{\pm 0.03}$	$0.19_{\pm 0.03}$	$0.1_{\pm 0.04}$	$0.23_{\pm 0.03}$	$0.19_{\pm 0.04}$
GA	$0.2_{\pm 0.08}$	$0.15_{\pm 0.05}$	$0.23_{\pm 0.06}$	$0.18_{\pm 0.08}$	$0.15_{\pm 0.06}$	$0.19_{\pm 0.05}$
QD init only	$0.16_{\pm 0.02}$	$0.2_{\pm 0.02}$	$0.25_{\pm 0.03}$	$0.16_{\pm 0.02}$	$0.24_{\pm 0.06}$	$0.22_{\pm 0.04}$
QD-quality	$0.27_{\pm 0.1}$	$0.26_{\pm 0.08}$	$0.38_{\pm 0.06}$	$0.25_{\pm 0.07}$	$0.33_{\pm 0.1}$	$0.32_{\pm 0.09}$
QD-diversity	$\mathbf{0.36}_{\pm 0.07}$	$\mathbf{0.4}_{\pm 0.07}$	$\mathbf{0.45}_{\pm 0.04}$	$\mathbf{0.41}_{\pm 0.05}$	$\mathbf{0.41}_{\pm 0.09}$	$\mathbf{0.4}_{\pm 0.04}$
			Formality/Maturity			
LLM sampling	$0.13_{\pm 0.02}$	$0.14_{\pm 0.01}$	$0.19_{\pm 0.03}$	$0.1_{\pm 0.05}$	$0.23_{\pm 0.04}$	$0.19_{\pm 0.03}$
GA	$0.15_{\pm 0.05}$	$0.13_{\pm 0.05}$	$0.18_{\pm 0.06}$	$0.17_{\pm 0.07}$	$0.17_{\pm 0.06}$	$0.18_{\pm 0.05}$
QD init only	$0.13_{\pm 0.01}$	$0.15_{\pm 0.03}$	$0.23_{\pm 0.03}$	$0.13_{\pm 0.04}$	$0.23_{\pm 0.03}$	$0.22_{\pm 0.03}$
QD-quality	$0.21_{\pm 0.04}$	$0.24_{\pm 0.05}$	$\mathbf{0.31}_{\pm 0.06}$	$0.23_{\pm 0.08}$	$\mathbf{0.3}_{\pm 0.09}$	$0.28_{\pm 0.08}$
QD-diversity	$\mathbf{0.25}_{\pm 0.04}$	$\mathbf{0.31}_{\pm 0.05}$	$0.29_{\pm 0.06}$	$\mathbf{0.31}_{\pm 0.06}$	$0.29_{\pm 0.06}$	$\mathbf{0.3}_{\pm 0.06}$

Table I.23: The Distinct-2 results of multiple samples for each task in different pairs of preferences on LLAMA-3.2-3B-INSTRUCT.

	textsum	data2text	commongen	qa	textcon	dialog
			Length/Formality			
LLM sampling	$0.4_{\pm 0.05}$	$0.46_{\pm 0.05}$	$0.54_{\pm 0.07}$	$0.23_{\pm 0.12}$	$0.63_{\pm 0.07}$	$0.52_{\pm 0.1}$
GA	$0.38_{\pm 0.13}$	$0.34_{\pm 0.12}$	$0.38_{\pm 0.12}$	$0.34_{\pm 0.13}$	$0.34_{\pm 0.14}$	$0.36_{\pm 0.11}$
QD init only	$0.49_{\pm 0.05}$	$0.61_{\pm 0.04}$	$0.69_{\pm 0.05}$	$0.47_{\pm 0.07}$	$0.65_{\pm 0.08}$	$0.63_{\pm 0.07}$
QD-quality	$0.57_{\pm 0.19}$	$0.54_{\pm 0.14}$	$0.69_{\pm 0.13}$	$0.55_{\pm 0.13}$	$0.7_{\pm 0.18}$	$0.67_{\pm 0.14}$
QD-diversity	$\mathbf{0.78}_{\pm 0.05}$	$\mathbf{0.81}_{\pm 0.05}$	$\mathbf{0.81}_{\pm 0.1}$	$\mathbf{0.79}_{\pm 0.07}$	$\mathbf{0.83}_{\pm 0.08}$	$\mathbf{0.8}_{\pm 0.09}$
			Length/Maturity			
LLM sampling	$0.4_{\pm 0.05}$	$0.45_{\pm 0.06}$	$0.54_{\pm 0.08}$	$0.23_{\pm 0.12}$	$0.61_{\pm 0.07}$	$0.53_{\pm 0.09}$
GA	$0.31_{\pm 0.11}$	$0.26_{\pm 0.1}$	$0.37_{\pm 0.12}$	$0.28_{\pm 0.14}$	$0.24_{\pm 0.1}$	$0.27_{\pm 0.09}$
QD init only	$0.49_{\pm 0.05}$	$0.61_{\pm 0.04}$	$0.65_{\pm 0.06}$	$0.42_{\pm 0.07}$	$0.65_{\pm 0.08}$	$0.6_{\pm 0.09}$
QD-quality	$0.45_{\pm 0.17}$	$0.49_{\pm 0.13}$	$0.7_{\pm 0.11}$	$0.45_{\pm 0.12}$	$0.59_{\pm 0.18}$	$0.54_{\pm 0.19}$
QD-diversity	$\mathbf{0.73}_{\pm 0.11}$	$\mathbf{0.76}_{\pm 0.12}$	$\mathbf{0.84}_{\pm 0.05}$	$\mathbf{0.79}_{\pm 0.08}$	$\mathbf{0.82}_{\pm 0.08}$	$\mathbf{0.79}_{\pm 0.07}$
			Formality/Maturity			
LLM sampling	$0.4_{\pm 0.05}$	$0.47_{\pm 0.03}$	$0.54_{\pm 0.07}$	$0.23_{\pm 0.12}$	$0.62_{\pm 0.07}$	$0.53_{\pm 0.09}$
GA	$0.23_{\pm 0.1}$	$0.23_{\pm 0.1}$	$0.29_{\pm 0.11}$	$0.25_{\pm 0.12}$	$0.26_{\pm 0.11}$	$0.25_{\pm 0.09}$
QD init only	$0.48_{\pm 0.03}$	$0.51_{\pm 0.09}$	$\mathbf{0.66}_{\pm 0.05}$	$0.33_{\pm 0.12}$	$\mathbf{0.67}_{\pm 0.05}$	$\mathbf{0.6}_{\pm 0.07}$
QD-quality	$0.38_{\pm 0.1}$	$0.47_{\pm 0.13}$	$0.58_{\pm 0.13}$	$0.39_{\pm 0.14}$	$0.54_{\pm 0.17}$	$0.45_{\pm 0.15}$
QD-diversity	$\mathbf{0.55}_{\pm 0.08}$	$\mathbf{0.66}_{\pm 0.09}$	$0.53_{\pm 0.13}$	$\mathbf{0.57}_{\pm 0.12}$	$0.58_{\pm 0.12}$	$0.52_{\pm 0.11}$

References

[1] Herbie Bradley, Andrew Dai, Hannah Benita Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth O. Stanley, Grégory Schott, and Joel Lehman. 2024. Quality-Diversity through AI Feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

[2] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 1693–1701.

[3] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Trans. Assoc. Comput. Linguistics* 6 (2018), 317–328.

[4] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, 986–995.

[5] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1823–1840.

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). The Association for Computer Linguistics, 142–150.

[7] Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1173–1186.

[8] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`