Quantitative Analysis of Rhyme and Metre in LLM-generated Translations of Poetry

Jan-Felix Klumpp
Department of Linguistics
University of Tübingen
jan-felix.klumpp@uni-tuebingen.de

Abstract

While machine translation systems have been applied to many tasks with remarkable success, machine poetry translation has remained a challenge. This study investigates the capabilities of generative Large Language Models (LLMs) in the translation of poetry (taking Shakespeare's 154 sonnets as an example) from English to German. For this purpose, I define metrics that assess the reproduction of the rhyme scheme and the metre of the original in a quantitative way. The results indicate that LLMs still lag behind professional human translators (especially with regard to the reproduction of the rhyme scheme), but that their performance is significantly influenced by the chosen prompt strategy. In particular, iteratively refining the result emerges as a successful strategy in terms of the reproduction of the form, but this comes at the expense of other aspects such as grammaticality and the reproduction of the meaning.

1 Introduction

Poetry translation is a particularly difficult type of translation; it has even been claimed to be "one of the hardest possible tasks that can be considered in computational linguistics" (Genzel et al., 2010, p.1). The main reason for this is that the translation of a poem has to fulfill very different and sometimes conflicting expectations. One such conflict is that the translation has to resemble a poem in another language (in terms of content and style), but also constitutes a work of art in itself, which has led to the distinction between "free" vs. "faithful," also called "domesticated" vs. "foreignized" translations (Schleiermacher and Bernofsky, 2021, Venuti, 2017, cited in Jones and Irvine, 2013). Another conflict is that the original poem has not only a specific meaning, but also a specific poetic form, whose reproduction imposes boundaries on rendering the meaning faithfully (Genzel et al., 2010). Which of these aspects are deemed more or less important depends not only on the preferences of translators and readers, but also on aspects such as the author of the original (Gibbons, 1985), the target audience, and the context in which the translation is to be used (such that, e.g., a poem has to be translated differently if the translation is intended to be sung, Low, 2003).

Within just a few years, generative Large Language Models (LLMs) have permeated all areas of NLP applications. In the area of machine poetry translation, the flexibility of instruction-tuned LLMs to react to different prompt formulations may allow to produce translations more closely aligned to reader preferences. However, despite their impressive performance on many other NLP tasks, poetry translation still remains challenging for modern generative LLMs (Resende and Hadley, 2024).

The wide range of sometimes conflicting expectations also has consequences for the evaluation of translated poetry: A translation of a poem cannot be evaluated globally, but only with regard to specific aspects, which have to be defined clearly. So far, however, only little attention has been paid to analyzing individual aspects of LLM-translated poetry in a quantitative fashion. This study attempts to narrow this gap by conducting an analysis of LLM-generated translations of all 154 of Shakespeare's sonnets into German, using numerical metrics that can transparently be mapped to well-defined aspects of poetry translation. This automated yet interpretable way of evaluation allows to compare different LLMs and prompting strategies on a relatively large corpus of poetry.

A quantitative analysis of LLM-generated translations of poetry requires focusing on aspects that can reliably be characterized numerically. This study focuses on three such aspects, namely, the *reproduction of the meaning* of the original (understood as the rather literal semantic content of the poem), the *reproduction of the metre*, and the *repro-*

Proceedings of SIGHUM 2026, pages 189–201

duction of the rhyme scheme of the original. The reproduction of meaning and form (with the latter including metre and rhyme scheme) is often characterized as a tradeoff relation (Genzel et al., 2010), which makes these aspects particularly suitable for contrasting prompt strategies.[1]

2 Related Literature

Research on the topic of machine poetry translation comes from two somewhat different directions. One strand of research approaches the task mainly as a technical problem and focuses on developing systems that reach a better overall performance. Since the 2010s, several approaches to neural machine poetry translation have been explored: Genzel et al. (2010) and Ghazvininejad et al. (2018) use decoding constraints that restrict the options to those fulfilling some manually defined formal criteria (e.g., rhyming words). Chakrabarty et al. (2021) fine-tune multilingual LLMs on translated poetry to improve their performance. To overcome the lack of large training corpora of translated poetry, Song et al. (2023) and Huynh and Bao (2024) employ a style injection paradigm, where the original poem is first translated into prose and then converted into poetry by a separate style injection module, which can be trained using target language poetry and synthetic prose equivalents generated via backtranslation. Finally, multi-step prompting strategies are used, where additional information retrieved from the language model itself (Wang et al., 2024) or a database (Chen et al., 2025) is given back to the model to improve the quality of the translation.

While the aforementioned literature has focused on developing systems that can translate poetry better, the availability of modern generative LLMs that are able to perform a variety of different tasks has led to another strand of research, which investigates the properties of translations generated by these LLMs. Using a combination of automated metrics and manual evaluation, Resende and Hadley (2024) analyse LLM-generated translations of four poems according to criteria such as the content, the form and the style (in a Spanish/Portuguese-to-English setting). In a similar vein, Karaban and Karaban (2024) compare LLM-generated English translations of twelve poems by the Ukrainian poet Ivan Franko to those written by a human translator. Gao et al. (2024) compare translations of Chinese poetry by ChatGPT to those generated by conventional machine translation systems (DeepL and Google Translate). Other work that mostly uses manual close reading rather than automated quantitative metrics has investigated LLM-generated translations of individual poems from and into other (both high-resource and low-resource) languages, e.g., Kazakh to English (Manapbayeva et al., 2024), Arabic dialects to English (AlAfnan and Alshakhs, 2025), Spanish to Filipino and Spanish to English (Liwanag et al., 2024), and English, French and Arabic to each other (Almaktary, 2025). However, most work that investigates specific aspects of translated poetry does so only on a small number of poems, which adversely affects a meaningful statistical analysis. Using quantitative yet transparently interpretable metrics, this study extends the analysis of LLM-generated translations of poetry to a larger number of poems and experimental conditions.

3 Data and Methods

3.1 Data

The original texts to be translated are William Shakespeare's 154 sonnets, taken from a 1923 edition with modernized orthography, but without any other changes (Shakespeare, 2019). The LLM-generated translations are compared to five human-written German translations from the 19th and early 20th century: Dorothea Tieck's translation (probably from 1825, published posthumously) (Jansohn, 1992), Gottlob Regis' 1836 translation (Regis, 2004), Emil Wagner/Ludwig Walesrode's 1840 translation (Walesrode, 2007), Max Joseph Wolff's 1903 translation (Wolff, 1924), and Stefan George's 1909 translation (George, 2004).

3.2 LLMs

Five LLMs are evaluated, three decoder-only Transformer models (OpenAI's GPT-4o, Hurst et al., 2024, Anthropic's Claude 3.5 Sonnet, Anthropic, 2024a, Anthropic, 2024b, and Google's Gemini 1.5 Gemini Team, Google, 2024), one reasoning model (OpenAI's OpenAI o4-mini, OpenAI, 2025b), and one unified model that flexibly uses both strategies (OpenAI's GPT-5, OpenAI, 2025a).[2]

[1]The code and other materials related to this study can be found at: `https://github.com/janfelixklumpp/AnalyseLLMPoetryTranslation`.

[2]The API was used for `gpt-5` and the analysis conditions of `gpt-4o`, and the web user interfaces (accessed in late 2024 and 2025) were used otherwise.

Beside human-written translations, the LLM-generated translations are also compared to translations produced by DeepL (DeepL SE, 2017), a standard machine translation tool based on a neural network architecture.

3.3 Prompting Strategies

I use zero-shot prompting with different instructions, both in single-step strategies that directly generate the translation and in multi-step strategies, where the output is reused in subsequent steps. Three single-step prompting strategies (which center different aspects) are evaluated for all LLMs:

- Plain: a simple instruction to translate a poem from English to German,
- PlainForm: like Plain, but with an explicit instruction to reproduce the form of the original,
- PlainMeaning: like Plain, but with an explicit instruction to reproduce the meaning of the original as closely as possible.

For GPT-4o, I evaluate additional multi-step strategies, which can be divided into two categories: First, following Chen et al. (2024), I test iterative strategies, where the later steps are used to improve on a previous translation attempt:

- IterativeForm: the output from PlainForm is given back to the LLM together with the original poem and the instruction to reproduce the form better,
- IterativeMeaning: the output from PlainMeaning is used likewise, but with the instruction to reproduce the meaning better,
- IterativeMixed: the output from PlainMeaning is used, but the instruction is to reproduce the form better,
- IterForm2Steps: the procedure used for IterativeForm is applied a second time to the output of that condition.

Second, I manually decompose the translation into a two-step procedure, where the LLM is instructed to analyse the form of the poem first and then to translate it in a second step. This can be regarded as a task-specific instance of the Generated Knowledge Prompting strategy introduced by Liu et al. (2022).

- AnalysisTranslate: in a first step, the form of the original is analysed, and this information is then provided to the LLM together with the original and the instruction to produce a translation that reproduces the form,
- AnalysisRewrite: in a first step, the form of the original is analysed; then, the output from PlainMeaning is adjusted to that form.

The exact prompts can be found in Appendix A.

3.4 Evaluation

3.4.1 Reproduction of form

Shakespeare's sonnets almost all follow the prototypical form of the English (or Shakespearean) Sonnet, which is defined by its *metre* and its *rhyme scheme*. The metre is —for most English and German poetry— an abstract pattern of prototypically stressed and unstressed syllables in a line. The rhyme scheme characterizes which pairs of lines rhyme. In this study, a pair of lines is considered to rhyme if and only if the sequence of phonemes in one line from the last stressed vowel onward is also found at the end of the other line and the preceding consonant onset is different; however, in accordance with the conventions of much of German poetry (Peust, 2014), front rounded and front unrounded vowels are considered equivalent, as are long and short vowels.

These two aspects of the form are evaluated by comparing the translation to a reference metre and reference rhyme scheme. The reference metre for (almost all of) Shakespeare's sonnets is the iambic pentameter x/x/x/x/x/(x) (where "x" stands for an unstressed and "/" for a stressed position)[3], and their reference rhyme scheme is ABABCDCDEFEFGG (where all lines with the same letter rhyme with each other)[4]. Using the Phonemizer library for Python (Bernard and Titeux, 2021), I extract the sequence of stressed and unstressed syllables as well as the phoneme sequences at the end of each line from the translation. I define the following metrics to compare the phonetic properties of the translation to the reference metre and rhyme scheme:

1. The *metrical distance* is a line-wise weighted edit-distance (based on the Levenshtein distance, Levenshtein et al., 1966) between the

[3] except for Sonnet 145, where it is x/x/x/x/(x)

[4] except for Sonnet 99, where it is ABABACDCDEFEFGG, and Sonnet 126, where it is AABBCCDDEEFF

sequence of stresses in the translation and the reference metre. It is a real number greater than or equal to 0 and is to be minimized. The costs for insertion, deletion and substitution operations are derived from stress values that are assigned to all syllables in both the translation and the reference metre. The cost of insertion or deletion is equal to the stress value of the inserted or deleted element, while the cost of substitution is the difference between the two stress values. Stressed syllables are assigned a stress value of 1.0, unstressed syllables are assigned 0.8, secondarily stressed syllables 0.9, and optional syllables 0.4. These stress values are chosen such that substituting a stressed for an unstressed syllable or vice versa is less costly than inserting or deleting a whole syllable. This corresponds to the fact that such substitutions are frequent even in Shakespeare's original sonnets, while the number of syllables per line of the original always follows the metrical pattern (Heldner, 2013). If the translation and the reference differ in the number of lines, the shortest lines of the text with more lines are treated as if they were completely inserted or deleted (i.e., the corresponding metrical distance is the sum of stress values). The total metrical distance is divided by the number of lines in the reference metre to obtain the average metrical distance per line. Pseudo-code for the metrical distance algorithm can be found in Appendix B.

2. The *rhyme scheme similarity* is the harmonic mean (F_1 score) of precision and recall with respect to the overlap in rhyming pairs of lines in the reference rhyme scheme and the translation. It is a real number between 0 and 1 that is to be maximized. Each line is assigned a number (starting from the first line), and rhyming pairs are represented as pairs of these numbers. Such an evaluation based on individual pairs of lines allows to capture the internal structure of parts of the poem better than a global comparison of the rhyme scheme (as used by Resende and Hadley, 2024).

3.4.2 Reproduction of meaning

Previous studies on machine poetry translation have compared the machine-produced translations to human-written reference translations, using metrics such as BLEU (e.g., Genzel et al., 2010, Resende and Hadley, 2024, Karaban and Karaban, 2024) or BERTScore (e.g., Song et al., 2023, Wang et al., 2024). However, human-written translations of poetry often consider or prioritize aspects other than the mere reproduction of the meaning and thus cannot be considered objective gold standard references for this purpose. For this reason, I use the CometKiwi metric (Rei et al., 2022), which is a quality-estimation metric that does not rely on reference translations but directly assigns a score based on the source text and the candidate translation.

4 Results

4.1 Plain prompting strategies

Fig. 1 shows the average CometKiwi scores, metrical distances per line and rhyme scheme similarities for the three plain conditions (Plain, PlainForm, PlainMeaning) for all five models used (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5, OpenAI o4-mini, GPT-5), as well as the same metrics for the DeepL-generated translation and the five human translations. These results allow the following observations: First, there is a clear tradeoff relation between the reproduction of meaning (as approximated by the CometKiwi score) and the reproduction of form (as described by metrical distance and rhyme scheme similarity), with the DeepL translation prioritizing the reproduction of meaning, the human translations prioritizing the reproduction of form, and the LLM-generated translations in between. There is even some differentiation between human-written translations, such that, e.g., Gottlob Regis' translation, which conforms to the form of the original less strictly than most other translations (Horstmann, 2002), performs lower on the two form-related metrics. LLM translations also are more spread out along the axes (i.e., less consistent in the weight they assign to the criteria), as can be seen from their higher standard deviations (see Appendix C).

Second, the chosen prompting strategy strongly influences whether the reproduction of the form or of the meaning is prioritized. PlainForm improves the reproduction of the form for all five models. The differences for metre and rhyme between PlainForm and Plain are clearly significant under a significance level of 5% (see Table 1).[5] In contrast, PlainMeaning improves the reproduction

[5] All p-values in this study were obtained using paired t-tests.

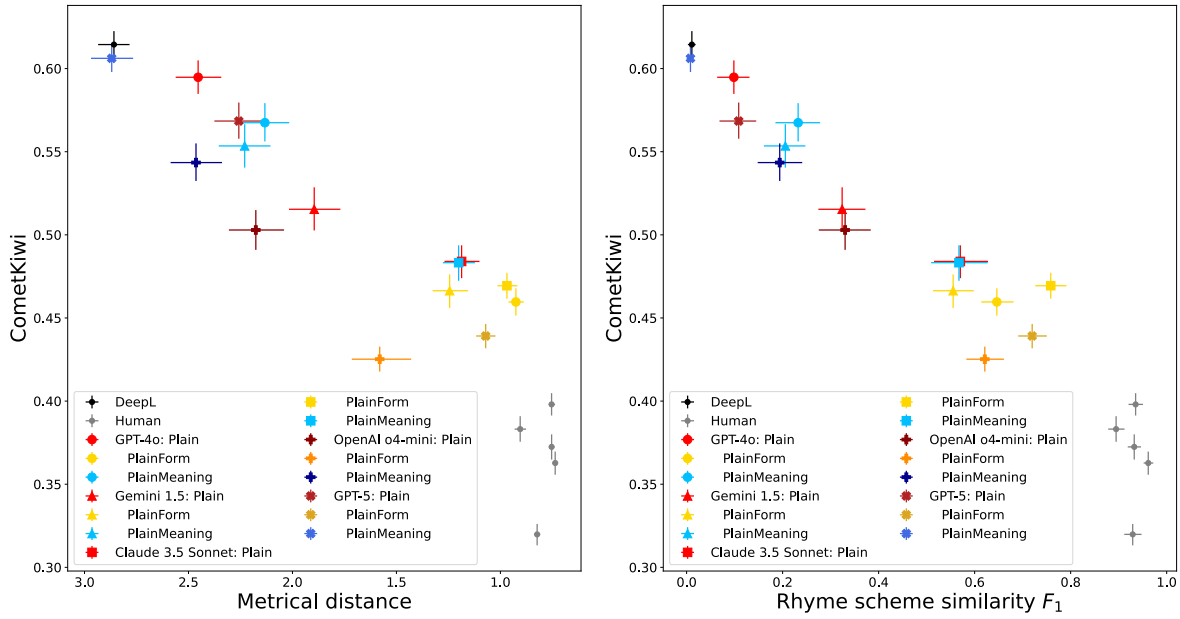

Figure 1: **Left**: Metrical distance per line (x-axis) and CometKiwi score (y-axis) for translations of Shakespeare's sonnets into German by several LLMs under single-step prompt strategies. **Right**: Rhyme scheme similarity (x-axis) and CometKiwi score (y-axis) for the same conditions. The markers indicate the average values, the error bars indicate 95% bootstrapped confidence intervals.

of meaning only for Gemini 1.5, OpenAI o4-mini, and GPT-5, while it has the opposite effect for GPT-4o. For all models (except possibly Claude 3.5 Sonnet) the prompt formulations that lead to a better reproduction of the form reproduce the meaning less well.

Among the LLMs, Claude 3.5 Sonnet on average reproduces the rhyme scheme best, and tends to reproduce the form to a considerable extent even when the prompt does not explicitly mention this aspect. In contrast, the other models' results are strongly prompt-dependent. OpenAI o4-mini on average performs worse than the other models (when only these three criteria are considered), with the CometKiwi scores being lower than those of otherwise comparable translations produced by other LLMs, particularly for PlainForm. This may be in part due to particularities of the resulting translations (see Section 4.3).

4.2 Multi-step prompting strategies

Fig. 2 shows the average CometKiwi scores, metrical distances per line and rhyme scheme similarities for the seven multi-step conditions (Iterative, IterativeForm, IterativeMeaning, IterativeMixed, IterForm2Steps, AnalysisTranslate, and AnalysisRewrite) for GPT-4o, as well as DeepL's translation, the human-written translations and the plain conditions for comparison.

Model	metrical distance	rhyme scheme similarity F_1
GPT-4o	1.36×10^{-58}	1.70×10^{-52}
Claude 3.5 Sonnet	5.01×10^{-7}	1.08×10^{-8}
Gemini 1.5	2.59×10^{-17}	5.91×10^{-13}
OpenAI o4-mini	1.02×10^{-9}	4.95×10^{-16}
GPT-5	3.02×10^{-45}	4.24×10^{-55}

Table 1: p-values for the hypothesis that the explicit inclusion of the target criteria in the prompt leads to improvement on form-related metrics for different models (i.e., PlainForm compared to Plain).

These results indicate that iterative prompting strategies, i.e., IterativeForm, IterativeMixed and IterForm2Steps, succeed in reproducing the form better than any of the plain conditions. The differences to PlainForm for metre and rhyme scheme are clearly significant under a significance level of 5% (see Table 2). With these iterative strategies, the metre is reproduced similarly well as in the human translations, while in terms of the rhyme scheme, no strategy reaches human-like performance. On the other hand, IterativeMeaning and IterativeMixed do not reproduce the meaning better than Plain or PlainMeaning. This mirrors the pattern observed for the plain conditions, where

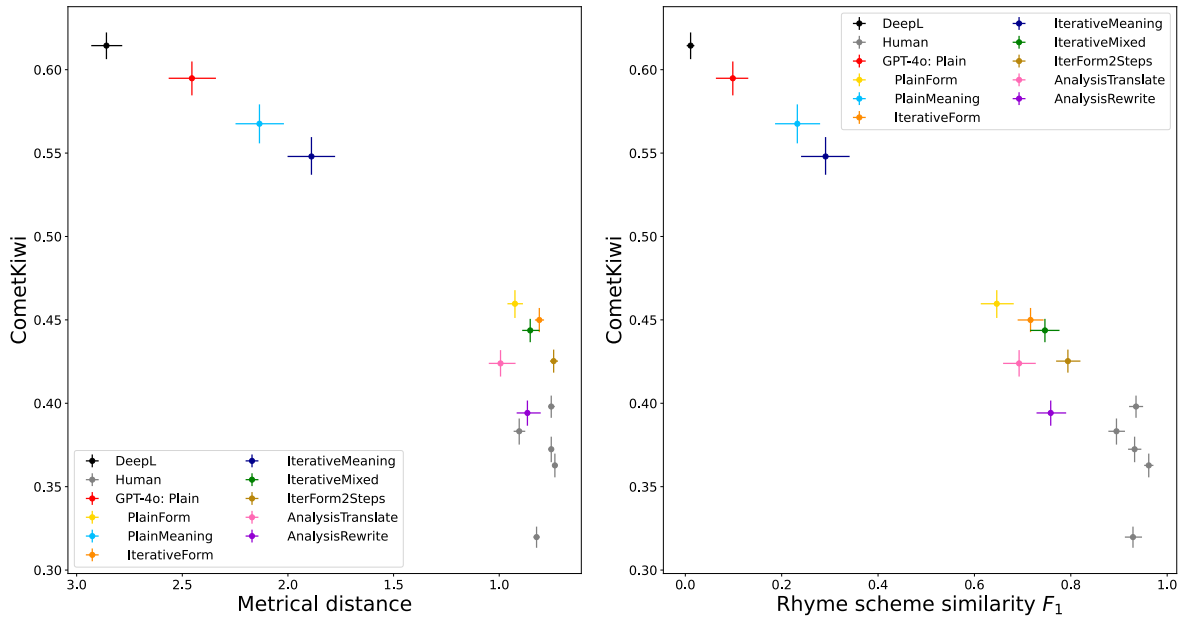

Figure 2: **Left**: Metrical distance per line (x-axis) and CometKiwi score (y-axis) for translations of Shakespeare's sonnets into German by GPT-4o under different prompt strategies. **Right**: Rhyme scheme similarity (x-axis) and CometKiwi score (y-axis) for the same conditions. The markers indicate the average values, the error bars indicate 95% bootstrapped confidence intervals.

Condition	metrical distance	rhyme scheme sim. F_1
IterativeForm	6.71×10^{-10}	0.00057
IterativeMixed	0.0058	0.000056
IterForm2Steps	6.93×10^{-8}	0.0000045

Table 2: p-values for the hypothesis that additional iterative steps lead to improvement on form-related metrics for GPT-4o (compared to PlainForm).

PlainForm clearly led to a better reproduction of the form, but PlainMeaning did not do so for the meaning.

The results for the analysis prompting strategies are less clear. Overall, AnalysisTranslate and PlainForm lead to similar performance on the form-related metrics. AnalysisTranslate however is associated with a lower average CometKiwi score. Furthermore, the generated outputs for other conditions without a separate analysis step often already started with a brief analysis (e.g., by mentioning that the poem is a sonnet), which means that adding a separate step often did not provide much further information.

AnalysisRewrite reproduces the form better than PlainForm (but not better than some of the iterative conditions), but has a lower CometKiwi score than any other prompt condition. The differences to PlainForm are significant under a 5% significance level, with $p = 0.041$ for metrical distance, $p = 6.98 \times 10^{-7}$ for rhyme scheme similarity, and $p = 1.50 \times 10^{-31}$ for CometKiwi.

4.3 Qualitative observations

The manual inspection of a few of the translated poems by the authors revealed some further particularities that do not fully manifest in the numerical metrics. Most importantly, those prompting strategies that lead to a closer adherence to the form of the original often also result in partially ungrammatical translations as well as divergent or even incomprehensible content (see Appendix D for an example). LLMs of the size of those used here do not usually make a lot of grammatical mistakes in text generation (neither in English nor German), nor were such mistakes produced at a high frequency in those translations that do not reproduce the form.

It is thus likely that the requirements of the form sometimes force the LLM to produce ungrammatical phrases. This would be an instance of a tradeoff relation not only between the reproduction of form and meaning, but also between poetic form and grammaticality. Furthermore, since the internal representations of an LLM are abstractions of patterns found in the training data, poetic licence (i.e., the fact that poets occasionally break grammatical

conventions in specific contexts or for a specific effect) in the training data may have contributed to the LLM learning patterns that allow deviations from regular grammar in poetry generally, even in cases or at a frequency that would be uncommon in human-written poetry. Finally, some of the grammar mistakes in the German translations can also be explained as interference from the English original.

Divergent or incomprehensible content cannot always be recognized from the CometKiwi score alone. Even in cases where it is still relatively high there may be completely incomprehensible or misleading lines. The reason might be that calculating one score for the whole poem obscures strong mismatches in the meaning of individual parts.

The qualitative analysis further revealed that an additional improvement step aiming at a closer reproduction of the form (as in IterativeForm, IterativeMixed, and IterForm2Steps) does usually not lead to whole parts of the previous attempt being copied. More often, a completely new translation is produced, which tends towards less literal word choices and a freer restructuring of the content (see Appendix D for an example).

Finally, LLM-produced translations sometimes follow a mostly consistent form that is not the one of the original (a strategy that is occasionally but rarely used by human translators as well). When this occurs, the alternative form usually employs metres and rhyme schemes common in German poetry (such as iambic hexameter and rhyming couplets). A notable exception are the reasoning model OpenAI o4-mini and the unified model GPT-5 (which depending on the prompt also uses reasoning) under the PlainForm strategy, which both produce a high number of identical rhymes (i.e., rhyme pairs where not only the phonemes from the nucleus of the last stressed syllable onward, but also that syllable's onset consonants are identical). Even though identical rhyme is rare in German poetry and considered a stylistic mistake in most contexts (Peust, 2014), almost half (72) of OpenAI o4-mini's and about two thirds (102) of GPT-5's translations feature at least two such rhyme pairs. 28 of OpenAI o4-mini's and 6 of GPT-5's translations even contain five or more pairs of identical rhymes (i.e., at least 10 out of 14 lines).

5 Discussion

5.1 Comparison to human translations

Compared to the human translators, the LLMs more often produced translations that match the form of the original only partially. In particular, reproducing the rhyme scheme remains challenging, and when it was reproduced, this often was accompagnied by a lower quality in terms of content reproduction, grammatical correctness, and style. On the other hand, the reproduction of the iambic pentameter was much less of a problem when the prompt clearly stated to reproduce the form.

Based on the CometKiwi scores alone, it may seem that on average the LLMs reproduce the meaning better than human translators. However, poetic meaning encompasses aspects other than the literal semantic content, and it is likely that the CometKiwi metric does not reflect all of its nuances equally well. Furthermore, the older and sometimes unconventional language of the human translations might have contributed to their lower CometKiwi scores. For example, Stefan George's translation (which performs worst on this metric) is characterized by a nonstandard lowercase-only orthography and idiosyncratic usage of punctuation.

5.2 Comparison between LLMs

All five models tested here (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5, OpenAI o4-mini, and GPT-5) showed the basic pattern of a tradeoff between the reproduction of form and meaning, and for all of them, the explicit instruction to reproduce the form, e.g., with the PlainForm prompt, led to a better performance on this criterion, and accordingly a less faithful reproduction of the meaning. While only three prompt formulations were tested for all models, these parallels in their behaviour suggest that different LLMs overall react to changes in the prompt similarly.

While the overall performance of all three standard LLMs (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5), as well as the unified GPT-5, was similar, the reasoning model OpenAI o4-mini performed worse, at least in terms of the metrics used here.

A particularity of translations produced by OpenAI o4-mini and GPT-5 was the frequent occurrence of identical rhyme, which seemingly stems from the reasoning process used by these models. Since identical rhyme is rare in German poetry and more difficult to use than "canonical" rhyme, this can neither be explained by a high frequency in the

training data nor by the need to reconcile the reproduction of the form with other aspects. Given that the correct form would be expected to be produced more easily than the erroneously produced one, this is probably not a problem of the general model capabilities. Rather, it seems that somewhere during the multi-step reasoning procedure, the goal shifts from reproducing the form of the original to producing a specific, entirely different form.

5.3 Effect of prompting strategies

Comparing the results of several prompting strategies revealed that they differ in their effectiveness to align translations to a certain target criterion (i.e., reproduction of the form or the meaning).

If the goal is to prioritize the reproduction of the form, it improves the performance of all tested LLMs a lot if this is clearly stated in the prompt. This only holds for some models when there should be a focus on the meaning, or at least it does not consistently manifest in the resulting CometKiwi score. One explanation is that human translators —from whose work the LLM learns the relevant distinctions during training— usually consciously decide whether they want to reproduce the form or not, and sometimes explicitly state that they do so. In contrast, they do not usually make such a binary decision for the reproduction of the meaning.

To improve the reproduction of the form, it is beneficial to give the translation attempt back to the LLM and to instruct it to produce a better translation. On average, doing so more than one time further improves the reproduction of the form (compare IterForm2Steps to IterativeForm), but further research would be necessary to systematically inquire how exactly the properties of the input translation affect the output. In contrast, dividing the task into separate steps did not consistently lead to further improvement. These observations align with Wu et al. (2025)'s findings for a general, non-poetic translation task, where iterative refinement led to a better performance than task decomposition too. It may be the case that an explicit analysis would be more relevant for less common poetic forms. On the other hand, however, the quality of the analysis might be lower for rare poetic forms, as language models recognize more frequent forms (such as the sonnet) more reliably (Walsh et al., 2024).

Not including the original poem in the final translation step (in AnalysisRewrite) strongly impairs the reproduction of the meaning, with the associated CometKiwi score being on average lower than in any other condition. This highlights the risks of splitting the process of poetry translation into two fully disconnected steps (literal translation and style injection). One of the challenges of poetry translation is that while it requires taking into account different aspects, the complex ways in which these aspects interact with each other make it difficult to divide the task into independent steps: Every part of the poem (be it a phrase, a line, or a longer section) influences the form and the meaning at the same time, and both form and meaning result from the interaction of several parts of the poem. Providing the original poem at every step prevents the LLM from moving too far away from the original meaning. Since the reasoning model OpenAI o4-mini presumably also divides the problem into subtasks, its lower performance and peculiar output might similarly stem from the impossibility to separate the task into isolated steps.

In general, LLM-generated translations that match the form of the original rather well often contain incomprehensible lines or grammatical mistakes. This suggests that a similar tradeoff relation as between the reproduction of the form and the meaning also exists between the reproduction of the form and other aspects (such as the adherence to the grammatical norms of the target language).

6 Conclusion

By using numerical and easily interpretable metrics describing the adherence of a translation of a poem to a specific form, this study confirms that the reproduction of the form of an English poem in its German translation remains challenging even for very large generative language models. However, the performance with regard to this aspect is strongly influenced by the chosen prompt strategy. In particular, giving the output back to the LLM together with the instruction to reproduce the form better helps to approximate the metre and rhyme scheme of the original. On the other hand, it is not possible to divide the task into fully independent steps, and the translations are on average better when the original poem is provided at each step of the translation process. These results might also be relevant for other domains where the output has to fulfill several criteria that interact in a complex way at the same time.

The results also confirm that there is a tradeoff

between different criteria. For both human translators and LLMs, reproducing the form better comes at the cost of having to compromise on other aspects (such as meaning, style and grammar). A translation that fulfills all criteria perfectly is impossible, but the influence of the chosen prompt strategy on the output suggests that —given further improvement with regard to the overall translation quality— LLMs can be used to tailor poetic translations to reader preferences more flexibly.

Limitations

While the restriction to a set of well-known poems that have often been translated into German allows a comparison to not only one but several human-written translations, it is likely that both the original poems and German translations are included in the training set of the LLMs, which may skew the results.

Furthermore, comparing the form of an original poem and its translation works best when the basic properties of this form are known to both the source and the target language tradition. In this case, the fact that the sonnet and the accentual-syllabic iambic pentameter are common in both English and German poetry allows a direct comparison, but further considerations about the transferability of the form as well as adjustments of the metrics could be necessary in situations where the form is specific to one of the languages or is defined differently in both of them.

More generally, this study considers only one author (William Shakespeare), one poetic form (the sonnet) and one language pair (English to German), and it is not certain to what extent the results hold for other kinds of poetry. In addition, the language of Shakespeare's sonnets (i.e., Early Modern English) differs from modern English, thus metrics such as CometKiwi may be less reliable here than they are for more recent texts.

Finally, the models tested here are closed-source models, which means that some of their exact specifications and settings are not known.

Acknowledgements

I would like to thank Michael Franke and Angelika Zirker as well as the anonymous reviewers for helpful feedback and advice. This work was supported by the Volkswagen Foundation through a Momentum grant.

References

Mohammad Awad AlAfnan and Thana Alshakhs. 2025. Bridging linguistic and cultural nuances: A comparative study of human and AI translations of Arabic dialect poetry. *Advances in Artificial Intelligence and Machine Learning*, 5(1):186.

Hussein Almaktary. 2025. Reflections on translating poetry through ChatGPT: Insights from Arabic, English, and French. *Journal of English Studies in Arabia Felix*, 4(1):1–12.

AI Anthropic. 2024a. The Claude 3 model family: Opus, Sonnet, Haiku. *Claude-3 Model Card*, 1:1.

AI Anthropic. 2024b. Claude 3.5 Sonnet model card addendum. *Claude-3.5 Model Card*, 3(6).

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in Python. *Journal of Open Source Software*, 6(68):3958.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

DeepL SE. 2017. Deepl translate.

Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai. 2024. Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1):1–10.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "poetic" statistical machine translation: Rhyme and meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Stefan George. 2004. *Shakespeare. Sonette*. Zeno.org. (Original work published in 1909).

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71, New Orleans, Louisiana. Association for Computational Linguistics.

Reginald Gibbons. 1985. Poetic form and the translator. *Critical inquiry*, 11(4):654–671.

Christina Heldner. 2013. On rhythm in poetry translation: A new Swedish translation of Shakespeare's sonnets.

Gesa Horstmann. 2002. *Shakespeares Sonette in Deutschland. Zur Geschichte der Übersetzungen zwischen dem 18. Jahrhundert und den Übertragungen von Stefan George und Karl Kraus*. Doctoral dissertation, Technische Universität Berlin. Available at https://webdoc.sub.gwdg.de/ebook/diss/2003/tu-berlin/diss/1995/horstmann_gesa.pdf.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Triet Minh Huynh and Quan Le Bao. 2024. Vietnamese poem generation & the prospect of cross-language poem-to-poem translation. *arXiv preprint arXiv:2401.01078*.

Christa Jansohn, editor. 1992. *Shakespeares Sonette in der Übersetzung Dorothea Tiecks, kritisch herausgegeben von Christa Jansohn*. A. Francke Verlag Tübingen.

Ruth Jones and Ann Irvine. 2013. The (un) faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101.

Viacheslav Karaban and Anna Karaban. 2024. AI-translated poetry: Ivan Franko's poems in GPT-3.5-driven machine and human-produced translations. In *Forum for Linguistic Studies*, volume 6.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Lois Mauri Anne Liwanag, Gabriel Luis Liwanag, and Leslie Anne Liwanag. 2024. AI in anthem: A comparative analysis of the English and Filipino ChatGPT 4 translations from the existing translations of the Philippine national anthem. *Recoletos Multidisciplinary Research Journal*, 12(2):91–102.

Peter Low. 2003. Translating poetic songs: An attempt at a functional account of strategies. *Target. International journal of translation studies*, 15(1):91–110.

Zhannura Manapbayeva, Gulkhadisha Zaurbekova, Karlygash Ayazbekova, Aigul Kazezova, and Kunsulu Pirmanova. 2024. AI in literary translation: ChatGPT-4 vs. professional human translation of Abai's poem 'Spring'. *Procedia Computer Science*, 251:526–531.

OpenAI. 2025a. GPT-5 system card.

OpenAI. 2025b. Introducing OpenAI o3 and o4-mini.

Carsten Peust. 2014. Towards a typology of poetic rhyme. In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic linguistics in typological perspective*, volume 55, pages 341–385. Walter de Gruyter GmbH & Co KG.

Gottlob Regis. 2004. *Shakespeare, William. Poetische Werke: Sonette*. Zeno.org. (Original work published in 1836).

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Natália Resende and James Hadley. 2024. The translator's canvas: Using LLMs to enhance poetry translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 178–189, Chicago, USA. Association for Machine Translation in the Americas.

Friedrich Schleiermacher and Susan Bernofsky. 2021. On the different methods of translating. In *The translation studies reader*, pages 51–71. Routledge.

William Shakespeare. 2019. *Shakespeare's Sonnets*. Wikisource. (Original work published in 1923).

Wai Lei Song, Haoyun Xu, Derek F. Wong, Runzhe Zhan, Lidia S. Chao, and Shanshan Wang. 2023. Towards zero-shot multilingual poetry translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 324–335, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Lawrence Venuti. 2017. *The translator's invisibility: A history of translation*. Routledge.

Ludwig Reinhold Walesrode. 2007. *William Shakspeare's sämmtliche Gedichte.* Wikisource. (Original work published in 1840).

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? Poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603, Miami, Florida, USA. Association for Computational Linguistics.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Max J Wolff, editor. 1924. *William Shakespeare: Sonette*. Projekt Gutenberg-DE. (Original work published in 1925).

Di Wu, Seth Aycock, and Christof Monz. 2025. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

A Prompts

Plain:
Please translate the following poem to German:
<original poem>

PlainForm:
Please translate the following poem to German. Please make sure that your translation reproduces the form of the original (rhyme and meter):
<original poem>

PlainMeaning:
Please translate the following poem to German. Please make sure that your translation reproduces the meaning of the original as closely as possible:
<original poem>

IterativeForm:
You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.
Original:
<original poem>
Attempt at translation:
<translation from Condition B>

IterativeMeaning:
You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the meaning of the original better.
Original:
<original poem>
Attempt at translation:
<translation from Condition C>

IterativeMixed:
You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.
Original:
<original poem>
Attempt at translation:
<translation from Condition C>

IterForm2Steps:
You are provided with an English poem and an attempt at a German translation. Please suggest a translation to German that reproduces the form of the original (rhyme and meter) better.
Original:
<original poem>
Attempt at translation:
<translation from Condition B>

AnalysisTranslate, first step:
Please tell me the meter and rhyme of the following poem (in the format: "Meter: [meter]; Rhyme scheme: [rhyme scheme]").
Poem:
<original poem>

AnalysisTranslate, second step:
Please translate the poem below to German. Please make sure to reproduce the meter and rhyme scheme of the original, making use of the given additional information.
Information on meter and rhyme scheme:
<output of the first step>
Poem:
<original poem>

AnalysisRewrite, first step:
Please tell me the meter and rhyme of the following poem (in the format: "Meter: [meter]; Rhyme scheme: [rhyme scheme]").
Poem:
<original poem>

AnalysisRewrite, second step:
Please translate the following to German:
<output of the first step>
AnalysisRewrite, third step:
Bitte schreiben Sie den folgenden Text in ein Gedicht mit den folgenden Eigenschaften um:
<output of the second step>
Text:
<translation from Condition C>

B Pseudocode for the metrical distance algorithm

$metr\_dist = 0$;
while $no\_of\_lines(transl) > no\_of\_lines(ref)$
do
  $metr\_dist = metr\_dist +$
    $\min_x(sum\_of\_stress\_values(transl_x))$;
  from $transl$ remove
    $transl_{argmin_x(sum\_of\_stress\_values(transl_x))}$;
end
while $no\_of\_lines(ref) > no\_of\_lines(transl)$
do
  $metr\_dist = metr\_dist +$
    $\min_x(sum\_of\_stress\_values(ref_x))$;
  from ref remove
    $ref_{argmin_x(sum\_of\_stress\_values(ref_x))}$;
end
$metr\_dist = metr\_dist +$
  $\sum_{x \in 1,\ldots,no\_of\_lines(ref)}(ldist(transl_x, ref_x))$;
$metr\_dist = metr\_dist \div no\_of\_lines(ref)$

Algorithm 1: Metrical distance algorithm

$transl$ is the sequence of sequences (lines) of stress values extracted from the translation. ref is the sequence of reference line metres. $transl_x$ is the stress pattern of the x-th line of the translation, and ref_x is the reference line metre for the x-th line.

$ldist(translation, reference)$ is a standard edit distance with three operations (insertion, deletion, substitution), where the cost of insertion and deletion is the stress value and the cost of substitution is the difference in stress values.

C Means and standard deviations for all settings

Note: All results are given in the format *mean/standard deviation*.

	Translator, prompting strategy	CometKiwi
1	Dorothea Tieck	0.363/0.045
2	Gottlob Regis	0.383/0.049
3	Ludwig Walesrode	0.372/0.048
4	Max Joseph Wolff	0.398/0.042
5	Stefan George	0.320/0.041
6	DeepL	0.614/0.051
7	GPT-4o, Plain	0.595/0.064
8	GPT-4o, PlainForm	0.460/0.052
9	GPT-4o, PlainMeaning	0.568/0.074
10	GPT-4o, IterativeForm	0.450/0.046
11	GPT-4o, IterativeMeaning	0.548/0.071
12	GPT-4o, IterativeMixed	0.444/0.045
13	GPT-4o, IterForm2Steps	0.425/0.044
14	GPT-4o, AnalysisTranslate	0.424/0.050
15	GPT-4o, AnalysisRewrite	0.394/0.048
16	Claude 3.5 Sonnet, Plain	0.484/0.062
17	Claude 3.5 Sonnet, PlainForm	0.469/0.049
18	Claude 3.5 Sonnet, PlainMeaning	0.483/0.067
19	Gemini 1.5, Plain	0.515/0.082
20	Gemini 1.5, PlainForm	0.466/0.064
21	Gemini 1.5, PlainMeaning	0.553/0.083
22	OpenAI o4-mini, Plain	0.503/0.076
23	OpenAI o4-mini, PlainForm	0.425/0.048
24	OpenAI o4-mini, PlainMeaning	0.544/0.072
25	GPT-5, Plain	0.568/0.070
26	GPT-5, PlainForm	0.439/0.046
27	GPT-5, PlainMeaning	0.606/0.052

	Metrical distance	Rhyme scheme similarity F_1
1	0.74/0.077	0.96/0.062
2	0.91/0.176	0.89/0.108
3	0.75/0.083	0.93/0.087
4	0.75/0.096	0.94/0.094
5	0.82/0.091	0.93/0.113
6	2.86/0.473	0.01/0.050
7	2.45/0.696	0.10/0.211
8	0.93/0.235	0.65/0.215
9	2.13/0.736	0.23/0.294
10	0.81/0.134	0.72/0.171
11	1.89/0.712	0.29/0.325
12	0.85/0.258	0.75/0.195
13	0.74/0.119	0.79/0.158
14	0.99/0.403	0.69/0.216
15	0.87/0.365	0.76/0.192
16	1.19/0.529	0.57/0.356
17	0.97/0.303	0.76/0.206
18	1.20/0.489	0.57/0.377
19	1.90/0.772	0.32/0.312
20	1.24/0.540	0.56/0.272
21	2.23/0.795	0.21/0.274
22	2.18/0.842	0.33/0.343
23	1.58/0.898	0.62/0.244
24	2.46/0.787	0.19/0.294
25	2.26/0.755	0.11/0.243
26	1.07/0.290	0.72/0.186
27	2.87/0.639	0.01/0.044

D Example of refinement over several steps

The following translations of Shakespeare's Sonnet 29 by GPT-4o under the conditions PlainForm, IterativeForm, and IterForm2Steps serve as an example of the changes over several iterations. They are annotated for significant

content differences to the original as well as ungrammatical phrases. Red marks ungrammatical parts, yellow marks parts of the translations that strongly differ from the content of the original, and green marks changes in the order of contents.

William Shakespeare: Sonnet 29 (Shakespeare, 2019)
When, in disgrace with fortune and men's eyes,
I all alone beweep my outcast state,
And trouble deaf heaven with my bootless cries,
And look upon myself, and curse my fate,
Wishing me like to one more rich in hope,
Featur'd like him, like him with friends possess'd
Desiring this man's art, and that man's scope,
With what I most enjoy contented least;
Yet in these thoughts myself almost despising,
Haply I think on thee,—and then my state,
Like to the lark at break of day arising
From sullen earth, sings hymns at heaven's gate;
For thy sweet love remember'd such wealth brings
That then I scorn to change my state with kings.

Translation by GPT-4o, PlainForm
CometKiwi: 0.383, Metrical distance: 0.80, Rhyme scheme similarity F_1: 0.42

Wenn ich, von Glück und Menschens Blick verbannt,
follows English rather than German morphology
Ganz einsam meinem ausgestoß'nen Stand
Nachweine, taubem Himmel Schmerz gesandt,
Verfluche mich und meiner Sterne Hand,
Und wünschte, hoffnungsvoll wie jener Mann,
Ansehnlich wie der Freundesreiche dann,
the phrase is missing a verb
Begehrte dies' Talent und jenen Plan,
Und schätzte nicht, was ich doch lieben kann;
Doch mitten in solch schmachvollem Verlangen
Denke ich, plötzlich, glücklich nur an dich.
Dann steigt mein Herz, wie Lerchen aufgesprungen,
Die singen, wenn der Tag das Dunkel bricht;
"from sullen earth" and "at heaven's gate" is not reproduced in this translation
Denn deines süßen Denkens reiche Gaben
Lässt mich auf Königsglück von Herzen verzichten.
number agreement mistake

Translation by GPT-4o, IterativeForm
CometKiwi: 0.427, Metrical distance: 0.81, Rhyme scheme similarity F_1: 0.36

Wenn ich, verstoßen von des Schicksals Hand,
Ganz einsam weine meinen schalen Stand,
intransitive verb used with a direct object
Und tauben Himmel ruf mit leerem Schrei,
missing article
Mich anseh' selbst und meinem Schicksal groll',
Wünscht' ich, wie jener hoffnungsreich zu sein,
Gesellt wie er, mit Freunden wohl bedacht,
Begehrt' ich dies' Talent, des andern Schein,
Und das, was mir gefällt, schätzt' ich recht flach.
Doch mitten in solch fast verhasstem Trachten
Denk' ich an dich – und plötzlich strahlt mein Los;
Wie Lerche, die im Morgengrauen erwacht,
missing article
Die freudig singt, als sei die Nacht nicht groß.
"from sullen earth" and "at heaven's gate" is not reproduced in this translation
Denn süße Liebe, die mir dich beschert,
Macht, dass ich Königs Glück verschmäh' und wehr'.

Translation by GPT-4o, IterForm2Steps
CometKiwi: 0.438, Metrical distance: 0.66, Rhyme scheme similarity F_1: 0.92

Wenn ich, verachtet, fern von Glück und Gunst,
Allein bewein' mein ausgestoßnes Los,
Und ruf' zum tauben Himmel meine Kunst,
Verfluchend mich und mein Verhängnis bloß;
Wenn ich mich wünsch', wie jener reich an Mut,
Und schön wie der, beliebt und hochverehrt,
Begehr' dies' Mannes Kunst, des andern Gut,
Und was ich sonst geliebt, erscheint mir leer;
Doch denk' ich dann an dich in meinem Schmerz,
Da schwingt mein Geist sich auf in lichten Klang,
Gleich wie die Lerche, die mit frohem Herz
Vom düstren Acker steigt zum Morgensang.
"at heaven's gate" is not reproduced in this translation
Dein süßer Liebes gruß so reich mich macht,
Dass ich kein Königs königreich betracht'.

Note especially the third quartet (ll. 9-12) in the IterForm2Steps translation, where the contents of the original are compressed, expanded and reordered more freely than in the previous iterations to produce a more coherent and smoother translation. The ll. 9-12 can be translated back to English as "Yet if I think of you in my pain, / then my spirit arises into bright sound, / like the lark, who with a happy heart / rises from the dark fields for its morning song." Note also the less literal translation of words at the rhyming ends of lines, e.g., "Mut" (literally, "courage") for "hope", and "beliebt und hochverehrt" (literally, "popular and revered") for "with friends possess'd".