jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Interpretation, Argument, Evaluation

A Workflow for Assessing LLM-Generated Interpretations of Poetry

Axel Pichler[1]
Martin Endres[2]
Nils Reiter[3]

1. Department of German Studies, University of Vienna, Vienna, Austria.
2. Institute for German and Dutch Philology, Freie Universität Berlin, Berlin, Germany.
3. Department of Digital Humanities, University of Cologne, Cologne, Germany.

Citation
Axel Pichler, Martin Endres, and Nils Reiter (2026). "Interpretation, Argument, Evaluation. A Workflow for Assessing LLM-Generated Interpretations of Poetry". In: *Journal of Computational Literary Studies* 5 (1). 10.48694/jcls.4312

Date published 2026-01-14
Date accepted 2025-12-21
Date received 2025-06-11

Keywords
interpretation, Large Language Models, generation, evaluation

Reviewers
Anonymous Reviewer, Anonymous Reviewer

Note
This paper has been submitted to the journal-only track of JCLS.

Abstract. This paper examines how interpretations of poems generated by LLMs can be evaluated in a way that meets standards from literary studies. To this end, we develop and evaluate a workflow that draws on reference data from literary studies and their argumentative structures when generating interpretations. This enables the generation of interpretations that themselves exhibit such structures and can be evaluated with respect to both their argumentative coherence and literary scholarship standards. Our experiments demonstrate that this workflow can be applied successfully, and that the model under investigation generate reasonable descriptions of the poems, but fail at more abstract interpretative tasks.

1. Introduction

Despite the increasing diversification of literary studies in recent decades, the interpretation of literary texts remains one of its central practices. This is evident not only in the prominence of chapters on interpretation in key introductory works, but also in empirical studies highlighting the prevalence of interpretive articles in scholarly journals.[1] This focus on interpretation contrasts with current trends in computational literary studies, where machine learning methods such as large language models (LLMs) have been employed primarily for text analytic questions, which typically involve classification problems such as genre attribution or sentiment analysis.[2] Classification, the task of assigning previously defined categories to instances, can also be understood as a subform of description that is grounded in a theory and/or taxonomy of the relevant domain – for example, a theory of literary genres. Interpretation, by contrast, draws on specific theories of meaning – such as intentionalist or anti-intentionalist approaches – and, often in conjunction with text descriptions, attributes meanings to texts or text elements. An illustrative example may clarify this distinction: Identifying Hugo von Hofmannsthal's poem *Mein Garten* as a sonnet – on the basis of its adherence to the

1. For example, in Martus (2021, 48–54), which draws on a corpus linguistic study of the renowned journal *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* published in German-speaking countries, 630 of the 2,430 articles published there between 1923 and 2018 were identified as interpretations from the field of Modern German Literature.
2. For classification tasks see the overview in Bamman et al. (2024). For the few exceptions of studies that deal with the interpretive competence of LLMs, see section 2.

characteristic features of this formally well defined genre – relies on a genre theory but does not presuppose a particular theory of meaning. By contrast, claiming that the poem explores the opposition between art and nature presupposes a theory of meaning that enables the attribution of such semantic properties to the poem.

In this work, we take initial steps towards exploring how interpretations of literary texts generated by LLMs can be evaluated. Here, we refer specifically to instruction fine-tuned LLMs that operate according to the 'prompt-and-generate' paradigm[3], enabling them to generate coherent outputs in response to open-ended textual prompts.

Any attempt to explore the potential of LLMs for generating literary interpretations must contend with a foundational characteristic of literary studies: There are different conceptualizations of what it means to interpret a text. These differing conceptualizations are associated with distinct standards by which actual interpretations are assessed or evaluated. Such standards are often contested within literary theory and are frequently described as theory-dependent. Nevertheless, to evaluate generated interpretations of literary texts in a consistent and transparent manner, an explicitly formulated standard is needed – ideally one that is accepted independently of specific theoretical presuppositions. In other words, a well-defined set of evaluation metrics is necessary to enable the assessment of LLM-generated interpretations in the first place.

To explore such a set of metrics and address the issue of the theory-dependence of existing practices of interpretation and evaluation in literary studies, we propose a method of generating interpretations that reduces them to their argumentative core. By 'argumentative core', we refer to the fundamental argumentative structure that underpins a literary interpretation, independent of its stylistic or rhetorical presentation. While the modes of textualization in literary interpretations vary depending on the approach, it seems largely undisputed that they involve argumentation.

Building on this notion of an argumentative core, this paper seeks to identify an evaluative framework – drawn from theoretical debates – and to select and refine criteria derived from this framework to test whether they are suitable for assessing the argumentative core of LLM-generated interpretations. Our primary objective, therefore, lies in the selection and refinement of evaluation criteria and in demonstrating that they can be applied in an intersubjectively consistent manner. By contrast, the actual evaluation of LLM-generated interpretations falls outside the scope of this study. We contend that a meaningful evaluation becomes feasible only when such criteria are explicitly defined and their application ensures a high level of consistency among evaluators.

In addition, we would like to highlight several further limitations of this study in order to prevent potential misunderstandings. First, we do not address the question of how LLMs generate meaning – nor how this process differs from human meaning-making – and what implications this has for their alleged 'understanding of language'.[4] For heuristic purposes, we adopt what we refer to as a *pretense stance*: We treat LLM outputs as if they were produced by intentional agents, while fully acknowledging

3. For a description of this paradigm and its differences to alternative applications of LLMs, see Liu et al. (2023).
4. A central reference point in this ongoing debate is Bender et al. (2021), who argue that LLMs merely mimic interpersonal language use, ultimately only predicting the next word, and therefore – particularly from an intentionalist perspective – cannot be considered genuine producers of meaning.

that these models do not possess genuine mental states. This interpretive strategy allows for a pragmatically useful engagement with LLM-generated texts, particularly in communicative and evaluative contexts. Conceptually, this stance is grounded in an externalist view of meaning, which holds that meaning does not arise from internal mental representations but from social, pragmatic, and interpretive practices. Within this framework, linguistic outputs are treated as meaningful insofar as they can be situated within communicative contexts and interpreted through interaction. This perspective is compatible with a range of externalist positions in current debates on LLMs and meaning, including accounts following Dennett's intentional stance – which legitimizes mentalistic attributions based on their explanatory utility rather than ontological commitments – as well as accounts of derived intentionality such as those proposed in Borg (2025) or Koch (2025).[5]

A second limitation pertains to our theoretical orientation within literary studies. Just as there is no single literary theory, there is no monolithic discipline of literary studies, but rather a plurality of approaches. Our perspective is rooted in a specific tradition – namely, the German-language debates on literary theory informed by analytical philosophy: analytical literary theory.[6] This approach is not characterized by adherence to a specific method but is instead defined by its commitment to scientific standards, conceptual clarity, precise question formulation, and rigorous argumentation (Köppe 2008).

In summary, this paper introduces a workflow for generating and evaluating literary interpretations using large language models. We begin by outlining the theoretical background, offering a brief overview of key debates on interpretation within literary theory. Next, we present different evaluation models, from which we select one – the framework by Strube (1992) – and justify this selection. We then address the question of how literary interpretations – specifically of poetry – can be generated by instruction-following LLMs in a way that aligns with Strube's criteria. To achieve this, we adopt an approach based on the argumentative reconstruction of interpretations, ensuring that the generated texts can be systematically evaluated. Therefore, we operationalize Strube's criteria in detail. Subsequently, we also describe the construction of reference/training data based on the argumentative reconstruction of existing poem interpretations. This is followed by an outline of the experimental setup, the presentation of results, and finally, the conclusion, which includes a discussion of limitations and suggestions for future research.

2. Interpretation in Literary Theory and CLS

Interpretation is a central concept in literary studies.[7] The term is used to describe both the act of interpreting and the written results of this act, which can refer to either a single statement about a literary text or a complete essay dedicated to an exegesis.

5. An extensive and reflective justification of the added value of externalist positions in the debate on whether LLMs 'understand' is provided by Jannidis et al. (2025).
6. All direct quotations from German-language research are reproduced in English translation produced by DeepL.
7. An excellent overview of the current debate on interpretation in literary studies provides Descher et al. (2015), see also Davies and Matheson (2008).

However, the meaning of the term remains a subject of ongoing debate.[8] Bühler (1999), for example, describes 17 different uses of the word 'to interpret' with regard to the exegesis of texts in German. This ambiguity arises from the fact that the meaning of the term varies depending on its context of use and is influenced by related concepts – such as meaning, text, and work of art – as well as the theoretical frameworks in which these are embedded. Given this complexity, we refrain from proposing a single, fixed definition of interpretation. Instead, we adopt a scheme developed by Göran Hermerén, which provides a structured way to capture its diverse uses. Göran Hermerén describes 'interpretation' as the following relation between five variables: "X interprets Y as Z for U in order to V" (Herméren 1983, 142). This scheme makes it possible to differentiate between types of interpretation based on the definition of the variables: Depending on which object is interpreted in which way and with which purpose, a different type of interpretation results. According to Hermerén, the different types of interpretation correspond to different criteria to determine their 'correctness'.[9]

However, the correctness or truth of interpretation statements is only one of several criteria that can be used to assess interpretations. Literary theory has worked out numerous such criteria, which to some extent were and are always determined by their theoretical and theoretical-historical standpoint. In 1992, Werner Strube proposed a set of criteria for the assessment of interpretations based on the language use and dominant interpretational practices in literary studies.[10] Strube draws on the distinction between 'Auslegung' (exegesis) and 'Deutung' (interpretation) in German: He understands 'Auslegung' as the use of a specific scheme to interpret parts of a given text. 'Interpretation', on the other hand, refers to the combination of several such schemes into a final interpretation that refers to the entire text. Based on this distinction, Strube identifies four dimensions of the given practice of interpretation in literary studies and outlines relevant assessment criteria: 1) the way in which literary texts are described in literary studies, 2) the exegesis of a text, from which 3) the interpretation of a text differs, and 4) the mode of argumentation. For each dimension, he specifies conditions for their successful or unsuccessful realization. For the description in disciplinary terminology of literary studies, these are accuracy, relevance and appropriateness; for the exegesis, plausibility and historical coherence; for the interpretation, specificity, integrity and comprehensiveness; and for the argumentative structure, coherence, unforcedness and freedom from contradiction. It is controversial in literary theory whether such criteria

8. A widespread understanding, which goes back to Gilbert Ryle, is that interpretation is the attribution of meaning. However, as Axel Bühler (1999), among others, has shown, this definition only applies to the word or sentence level, but not to texts as a whole. With regard to the interpretative determination of texts that go beyond the sentence level, recent research speaks accordingly of "thematic interpretation", which is realized in statements of the following structure: "Text X is about y" (Winko et al. 2024, 166).

9. The question of whether there is one or more valid interpretations of individual literary phenomena or texts that contradict each other is a standing topic of debate in literary theory. The answers to this question range from interpretation-theoretical monism, which assumes that there is only one potentially correct interpretation within the framework of a particular type of interpretation, to interpretation-theoretical relativism. For an introduction to this debate: Davies and Matheson (2008); for a critique and rejection of interpretation-theoretical relativism in the sense of an acceptance of contradictory or incoherent interpretative statements: Descher (2017).

10. Alternative systematizations of evaluation criteria are offered by Beardsley (1981), Zabka (2008), and Descher et al. (2015, 47) as well as Petraschka and Descher (2019, 54–70). Winko et al. (2024, 495–516) systematizes the use of quality criteria in interpretative texts.

can be independent of the guiding theory and the overarching interpretative goals.[11] However, it should be noted that these debates primarily concern the application of the criteria to a complete interpretation and the practices associated with it, rather than to argumentative cores. Underlying these debates is the fundamental question whether there are assessment criteria that are valid *in general* and, if so, what these criteria might be.

The determination of generally valid criteria for assessing interpretations is closely tied to the problem of evaluating interpretations, or their arbitrariness. Lutz Danneberg has reconstructed this problem in the form of the following argument (Danneberg 1992, 15):

- If there are no acceptable (or justified) criteria for evaluating interpretations with respect to their validity claims, then interpretations cannot be be assessed in terms of their validity claims.
- If interpretations cannot be assessed in terms of their validity claims, then they are considered to be of equal evaluative rank.
- If interpretations are considered to be of equal evaluative rank, then their choice is arbitrary.

The question of the evaluation of interpretation is accordingly one of the central questions of literary theory.

Given the complexity of these issues and the central role of interpretation in literary studies, it may come as no surprise that, to the best of our knowledge, no attempts have yet been made to develop (quantitative) evaluation metrics for interpretations (generated by LLMs). However, there are first attempts to explore the range of possible criteria for such evaluations and the benchmarking of LLMs' text-interpretive abilities. One example is Jannidis et al. (2025): Their study investigates how well contemporary LLMs can "understand" poetry by probing nine core aspects of literary analysis – from meter and rhyme to figurative language and meaning – across increasing levels of interpretive complexity. The authors show that while LLMs perform well on semantic and interpretive tasks, they struggle with formally grounded operations such as scansion, phonetic pattern recognition, and culturally sensitive context integration. This study positions itself as an exploratory first engagement with the problem of interpretation, and thus as an attempt to delineate the literary- and communication-theoretical foundations for future benchmarking of LLM-generated interpretive statements. By contrast, in the present paper we develop and evaluate an approach that integrates the generation and the evaluation of interpretive statements so closely that a genuinely human – and ultimately quantifiable – assessment of them becomes already possible.

In addition to this, there are experiments from literature didactics that take a different approach by refraining from developing and explicating precisely defined evaluation

11. Paradigmatic for this is the following statement by Steffen Martus at the end of a contribution on the practice of interpretation in literary studies with regard to the multiple relationality of this practice to other factors: "Because […] the potential qualities of an interpretation are realized in varying degrees of quality and intensity, no schematically applicable evaluation rules can be given. 'Neither truth nor method guarantee […] that an interpretation is really good and acceptable to literary interpreters', and the question of 'how to distinguish between "good" and "not so good" practice' must be supplemented by the question: good for whom, for what and for which situation?" (Martus 2021, 74). Martus refers here to some of the variables which can also be found in Hermerén's definition above to characterize the type of interpretation and thus also its conditions for success. The quotations in the quote are taken from Hempfer (2018).

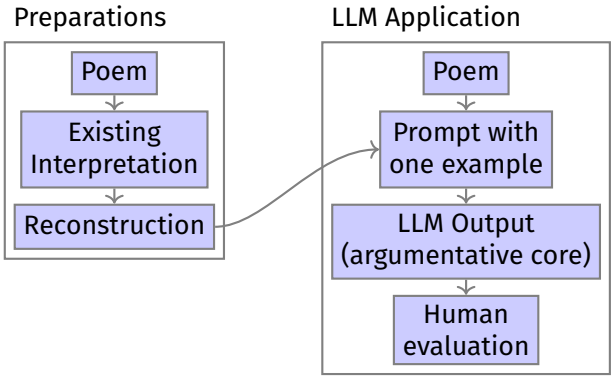

Figure 1: Schematic depiction of the workflow in this paper.

metrics. For example, Susteck and Perder (2023) use four canonical German poems to investigate the extent to which ChatGPT 3.5 can cope with writing tasks in high school poetry analysis. The authors found that ChatGPT performs very convincingly, particularly in the generation of interpretation hypotheses that "link texts with stereotyped, but often appropriate interpretation patterns due to their comparatively high degree of vagueness"(Susteck and Perder 2023, 12). Susteck and Perder's approach differs from the one presented by us in that 1) the evaluation of the generated texts is based on high school objectives – such as summary, classification within an epochal context, topic definition, and form analysis – rather than on explicitly operationalized evaluation criteria derived from a specific literary theory framework, 2) the OpenAI online chat interface is used and not the API, 3) no batch prompting is used, but a dialogical-chatbot-interaction, and, 4) interactive prompting was carried out.

3. Generating and Evaluating Interpretations

Our general workflow is depicted in Figure 1. Starting with a poem and existing interpretations, we first derive three argument reconstructions. These reconstructions are then used as examples in the prompt provided to the LLM, which also contains a new poem. The LLMs receive the reconstructed argumentations only in the form of individual statements, without any information about (a) which of Strube's levels they correspond to or (b) which argumentative function they fulfill.[12] The model then generates output in the same style (i.e., as the argumentative core of an interpretation). The generated outputs are subsequently evaluated through manual inspection.

In the following, we will explain this workflow in detail, discuss the rationale behind our choices, examine how it aligns with a specific form of evaluation, and justify why this particular form was selected.

3.1 Evaluation of Automatically Generated Interpretations

LLMs are capable of generating texts that resemble interpretations of poems. This requires nothing more than a prompt that, in addition to the request to interpret a poem, contains the poem itself.

12. In future experiments, this information will be provided in a modular fashion in order to determine which combination yields the most reliable results.

In principle, there are different ways on how to evaluate generated texts. A first possibility is the use of **evaluation metrics from Natural Language Processing:** In Natural Language Processing, the evaluation of generative language models is considered difficult.[13] Currently, metrics from the field of machine translation are often used. Here, the generated texts are compared with reference texts. These reference texts are the texts that the model should ideally generate. The two most common metrics for evaluating text generated by LLMs are BLEU (Papineni et al. 2001) and ROUGE (Lin 2004). BLEU calculates the n-gram overlap between the generated text and the reference text. While BLEU focuses on precision, ROUGE is oriented towards recall and distinguishes between different variants: ROUGE-N, which examines the n-gram overlaps, ROUGE-L, which examines the longest common subsequence instead of the n-gram overlap, and ROUGE-S, which focuses on so-called skip-bigrams.

These metrics are not suitable for our purposes, as they were developed for unstructured text rather than for argument-like structured interpretations, which we aim to evaluate. While it would technically be possible to serialize the reconstruction into a stream of tokens, the linguistic variability of such texts is likely to be quite high. As a result, it is entirely possible for perfectly congruent interpretation arguments to be expressed in different words, making these metrics inadequate for our needs.

Another possibility is to build on recent praxeological research. Praxeology understands interpretation as one of many practices in everyday literary studies and considers interpretive texts as manifestations of these practices (Martus and Spoerhase 2022). From this perspective, it would make sense to evaluate LLM-generated interpretations by involving literary studies scholars using the **scientific questionnaire method**.[14] Such an approach would have the added benefit of not only evaluating but also providing valuable insights into the guiding background assumptions of the discipline. However, the design of these questionnaires would ultimately rely on existing evaluation criteria. For this reason, we have chosen a third option for the present paper: the use of existing **criteria from literary theory/literary studies**.

As explained in section 2, there are catalogs of such criteria, but considering their significance, it is surprising how few of them actually exist. In the following, we will work with those of Werner Strube (1992). The reasons for this choice are as follows: Firstly, Strube claims to adopt a descriptive approach. His criteria were created on the basis of actual interpretation practice. Secondly, the argumentative structure of the interpretations plays a central role in his catalog of criteria, which makes them particularly suitable for application in the context given here. Thirdly, another advantage of applying Strube's criteria to the argumentative cores of interpretations is that the interpretative goals characteristic of interpretations, as described by Herméren, can be disregarded. Strube's criteria, in the version adapted by us in the following, do not require their specification – but they do allow for the inclusion of these goals within a modular extension of the cores, if needed. Fourthly, research has already indicated that his criteria are specific enough in relation to the actual practice of interpretation and

13. For an overview see: Celikyilmaz et al. (2021).
14. The scientific questionnaire method involves systematically collecting and analyzing self-reported verbal and numerical data from respondents about their experiences and behavior. This is done using a self-administered scientific questionnaire, which can be distributed in person, by mail, online, or via mobile devices. Key elements are the respondents, the questionnaire, and the context in which it is completed (Döring 2023, 393ff.).

that the guiding theoretical criteria are largely acceptable within literary studies as well as suitable for operationalization (Köppe and Winko 2011). Section 3.3 will be devoted to the latter.

3.2 Generation of Interpretations Suited for Evaluation

Literary interpretations consist of distinct components such as thesis statements, textual evidence, analytical reasoning, and contextualization. Evaluating such interpretations holistically risks conflating these components, making it difficult to determine which aspects of the interpretation meet the required standards and which do not. For this reason, it is necessary to isolate and evaluate individual components of the generated text in relation to specific criteria. With simple prompt-based generation, it is neither clear (a) which components of the output can or should be evaluated with regard to which criteria or, if such criteria exist, how the generated text should be broken down into its components so that these criteria can be applied to the corresponding components, nor (b) which literary-theoretical assumptions the LLM realizes during generation. To address these challenges, we adopt a procedure that already suggests a certain output structure via the prompt: the generation of argumentative cores of interpretations. This approach simplifies the isolation and evaluation of the individual components. From the perspective of actual interpretative practice in literary studies, this method may seem unconventional, as such practices typically do not adhere to rigid organizational or structural schemes for the interpretations they produce. Nevertheless, we believe that this limitation is outweighed by the possibilities to guide the generation in such a way that the output is structured to align with the expected levels of output components, thereby facilitating a systematic evaluation according to the selected criteria from literary studies.

From a machine learning perspective, it makes sense to enrich the prompts with such reference data in order to achieve the above-mentioned goals. In this case, these data should consist of existing interpretations of literary texts. To assess their influence on the generation process and to evaluate texts generated with their help, it is useful to extract their central components in a structured form. To achieve this, we draw on a common practice in dealing with and analyzing scholarly texts: the reconstruction of their central arguments.[15]

Such reconstructions of arguments necessarily go beyond the literal wording of the texts examined. Descher and Petraschka (2018) identify the following dimensions of argument-reconstruction to which this applies in particular: reformulations, the clarification of text elements that require interpretation, the addition of argumentation steps, the sequence of arguments and the choice of argumentation scheme. Accordingly, reconstructions of arguments are themselves highly interpretative.

The guiding principle in our reconstructions is a specific version of the principle of charity: the aim of reconstructing the strongest possible argument from the texts. To achieve better alignment with Strube's evaluation criteria, we base our reconstruction of the

15. According to Bowell et al. (2020, 144) the "goal of argument-reconstruction is to produce a clear and completely explicit statement of the argument that the arguer had in mind. The desired clarity and explicitness are achieved by putting all of the argument, and nothing but the argument, into standard form: this displays the argument's premises, intermediate conclusions and conclusion, and indicates the inferences between them.".

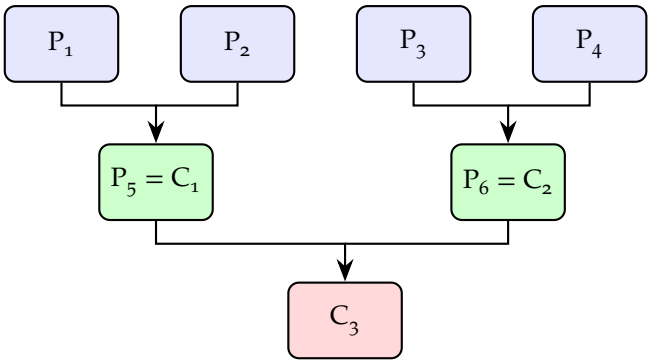

Figure 2: Reconstruction of an argument.

arguments on Strube's distinction between the three levels of interpretation: description, exegesis, and interpretation. Therefore we reconstruct only the first three levels of the potentially multi-level argumentation (Figure 2), whose lowest level consists of arguments, whose premises consist of textual descriptions (= P_1 - P_4), whose conclusions (= C_1 - C_2) themselves figure as premises (= P_5 - P_6) of the central argument of the interpretations (= C_3).

Reconstructions of the central arguments of an interpretation are not uncontroversial in literary studies. They are considered reductionist, as they ignore numerous aspects that can also be significant for the persuasive power of interpretations, such as their rhetorical and stylistic design (Albrecht and Danneberg 2021; Descher and Petraschka 2018). Reconstructions that focus on the core structure and theses therefore differ in several respects from those that are more closely aligned with the subject-specific culture and practice of interpretation in literary studies, such as those recently developed for the interpretation of canonical narrative texts in the German-speaking world as part of the ArguLit project at the University of Göttingen (Winko et al. 2024): While the latter strives for a "'dense description' of the argumentative contexts as well as the characteristics of the interpretative texts", taking into account "above all the diversity and linguistic complexity of the means of representation used" and accordingly reconstructs them on "a hermeneutic basis"(Winko et al. 2024, 43), the method used here is far more economical, focusing on the basic argumentative structure of the interpretative texts analyzed. Accordingly, we refer to the results of our argument reconstructions in the following as 'argument-like structured interpretation'.

It is important to note that such reconstructions serve only as proxies for the actual textual practices of literary studies, as they lack the detail and comprehensiveness of reconstructions that account for all dimensions of literary arguments. Nevertheless, we employ them for several pragmatic purposes. First, they allow literary interpretations to be distilled to their argumentative core, presenting them in a format that is both generic and comparable. This reduction makes them significantly easier to evaluate and contrast than the often rhetorically and stylistically complex source texts. Second, these reconstructions can be flexibly expanded with additional layers of information, such as the types of arguments employed, key textual passages, or meta-information, including the interpretation's objective or its underlying literary-theoretical framework. This makes them a modular and adaptable format for analyzing and comparing interpretive practices.

3.3 Operationalization of Werner Strube's Criteria for Evaluating an Interpretation

The criteria presented in Strube (1992) address different levels of the interpretive process: (1) the relationship between statements about the interpreted text and interpretive claims; (2) the relationship between these interpretive claims and higher-level interpretive statements; and (3) their argumentative connection. These criteria are expressed as single or multi-place predicates, which can be attributed either to individual statements within interpretations or to the relationships between them. In the following, we outline the stages of the interpretation process and specify the corresponding evaluation criteria. Each criterion is reformulated as a conditional rule, defining the conditions under which a given predicate may be attributed to (parts of) an interpretation. Where necessary, Strube's original formulations are adapted to fit our approach, which focuses primarily on the reconstruction of argumentative structures.

Descriptions are

- *empirically correct or accurate* if there is a correspondence between the description and the poem, i.e., if what the description claims is actually present in the text, and/or
- *appropriate* if the description serves as a premise for one of the arguments of the exegesis.

Exegeses of a text are

- *plausible*, if it "is sufficiently justified in the description assigned to it", that is, if the exegesis supported by the premises preceding it.[16]

Interpretations of a text are

- *integrative*, if the conclusions of all subarguments flow into the final argumentation as premises, and/or
- *specific enough*, if the conclusion of all sub-arguments is not vague and general, but instead makes statements about the subject that are as precise and detailed as possible.

The argumentation is

- *free of contradictions* if it does not contain any statements that are in logical opposition to each other, and/or
- *coherent or ordered* if "the exegesis is grounded in the description, the interpretation grounded in the exegesis" (Strube 1992, 198).

The three levels of evaluation correspond to the three levels of reconstruction with which we work (Figure 2). It follows from the frame of reference of the individual criteria that, evaluating an interpretation aimed at providing an overall interpretation of a text or

16. This is plausibility in the sense of justifiability, see Winko (2015).

poem, different criteria apply at different levels.[17] Criteria related to description apply exclusively to the premises of the reconstructed arguments; criteria relating to exegesis address the relationship between premises and material rules of inference at the second level; and criteria regarding interpretation pertain to the first or top level of the reconstructed argument. Strube himself, however, does not explicitly include material rules of inference in his list of criteria, likely because these are rarely made explicit in actual practice of interpretation. By material 'rules of inference', we refer to domain-specific principles of reasoning that connect premises and conclusions based on substantive knowledge.[18] Given their central importance for reconstructing the argument structure of interpretations, we have supplemented Strube's list by incorporating such explicit rules of inference. We reconstruct them as conditionals – that is, if-then sentences – whose components consist of the generalized premises or theses. These are considered *acceptable* if they are "collective convictions that have been accepted by the majority and/or by experts in the course of previous argumentation"(Winko et al. 2024, 41). It also seems reasonable to locate Strube's criterion of historical appropriateness here, which for him pertains to the so-called interpretive schemata. According to this criterion, statements are *historically coherent*, if they "correspond to what the author knew and could therefore have meant"(Strube 1992, 192).

Due to the fact that we work exclusively with reconstructed arguments, the following restrictions must be applied in the selection and use of Strube's criteria. We have omitted four of them: First, it is not possible to determine the *relevance* of the text-describing premises based solely on the reconstructions, as the literary-theoretical method is not explicitly mentioned within them.[19] This also renders the category of *comprehensiveness* obsolete, as this is dependent on the category of *relevance*. Secondly, it can be assumed that the premises supporting those arguments whose material rules of inference have been interpretatively inferred by us will be *appropriate*, since the construction of their rules of inference is based on precisely these premises. The same applies, thirdly, to the category of *integrity*: If the final material rule of inference is an inferred one, it is already reconstructed based on the aforementioned criterion. Fourth, we omit the category of *unforcedness*, as it conflicts with our understanding of the principle of charity: Our goal is to reconstruct the strongest possible arguments, which is why we exclude premises that cannot be integrated into the reconstruction. From a different perspective, this might appear as forced.

17. As Köppe and Winko (2011) critically note, Strube's system of categories is geared towards this case. Interpretations that pursue other goals that do not concern the entire literary text – e.g. the clarification of a poetic image or the intertextual context of a single verse – do not possess the third of Strube's levels of interpretation and would therefore be evaluated less favorably.
18. In using the term 'material rules of inference', we draw on the tradition established by Wilfrid Sellars (1953) and developed in contemporary inferentialism by Robert Brandom (2001), as well as the closely related notion of field-dependent warrants in StephenToulmin (2003)'s model of argumentation.
19. To asses the relevance of a description one would have to determine on the basis of the reconstruction, which literary-theoretical method could be involved in the interpretation, which leaves a great deal of room for interpretation and in many cases is not possible due to the lack of an explicit connection to literary theory or the use of vocabulary specific for a certain literary theory in many interpretations. This is at least the case in the poetry interpretations we have reconstructed. The situation is similar in the corpus of interpretative texts examined as part of the ArgLit project. Here, too, the literary-theoretical standpoint could only rarely be determined (Winko et al. 2024, 155).

4. Reconstructing Core Arguments

The experiments presented in this paper utilize reference data for the generation and evaluation of interpretations of poems. In the given case, reference data are interpretations of literary texts that are representative (not in a statistical sense) of the interpretation practices within the discipline. An interpretation can be considered representative if a) it is used in teaching or b) it is frequently cited. The former applies to texts from the Reclam publishing house. The Reclam Verlag is a German publishing house renowned for its pocket-sized editions of classic literature and accompanying interpretations. It plays a significant role in German education by making essential literary works accessible and affordable for students and readers.

At the end of the 1990s, Reclam published collections of interpretative texts on works by 12 canonical German-language poets as a series of collected interpretations of poems entitled *Gedichte und Interpretationen*. From this series, we have selected three interpretations: Jochen Schmidt's interpretation of Friedrich Hölderlin's *Hälfte des Lebens*(Schmidt 1984), Hans-Georg Kemper's interpretation of Georg Trakl's *Im Winter*(Kemper 1999) and Marco Meli's interpretation of Gottfried Benn's *Der Sänger*(Meli 1997). The guiding selection criteria were that (a) the poems analyzed should be described as comprehensively as possible in the context of the interpretation and (b) these descriptions should serve as premises in the actual interpretation.

To strengthen the influence of existing interpretations on the generation process and to establish a framework for the systematic evaluation of the generated texts, we employ argumentative reconstructions of these interpretations (see subsection 3.2). In reconstructing the arguments, we follow the recommended procedure in Brun and Hirsch Hadorn (2021) and supplement it with insights from Winko et al. (2024): The reconstruction process begins with a close reading of the text to be interpreted, followed by the development of a structured overview of the interpretation texts. Based on this overview, we identify the central thesis along with its supporting premises, reformulate unclear, incomplete, or inconsistent statements, and add missing premises or conclusions where necessary. In doing so, we balance two opposing principles: On the one hand, we aim to stay as close as possible to the original formulations; on the other, we seek to strengthen the reconstructed arguments to ensure they provide sound and coherent reasoning. This tension is particularly relevant when adding missing rules of inference.

Such domain-specific rules of inference are according to Winko et al. (2024, 263) "assumptions of interpreters [...] that underlie the plausibilization of their interpretative hypotheses, but usually represent the general framework assumptions or rules of the game for the plausibilization of interpretative hypotheses as implicit presuppositions that are potentially shared by many representatives of the subject". As implicit presuppositions, these rules are typically not explicitly formulated in the interpretation texts themselves and must therefore be supplemented in the reconstruction process.[20]

When adding inference rules, we proceeded as follows: After isolating the main thesis of an interpretation and identifying its supporting premises, we first examined which

20. In attempting to explicate such rules of inference, Winko et al. (2024, 269–272) have encountered numerous problems concerning, among other things, the degree of generality or the scope of these rules of inference.

inference rule could be supplemented with minimal intervention if no explicit rule was provided. In this process, we considered not only deductive arguments but also inductive reasoning and inference to the best explanation. For instance, if an interpretation argues – based on close readings – that the two stanzas of a poem stand in a relationship of allegory and reflection, we would reconstruct the argument as one from circumstantial evidence rather than explicitly formulating a deductive inference rule. Only when no inductive reconstruction was possible based on the given premises and conclusion we introduced a conditional inference rule, adhering to the principles of argumentation reconstruction outlined in Brun and Hirsch Hadorn (2021).

In the following we will use Hans-Georg Kemper's interpretation of Georg Trakl's poem *Im Winter* as an example to illustrate the procedure of argumentative reconstruction of an interpretation of a poem. Kemper's interpretation is divided into five parts, three of which focus on a particular dimension of the poem: After a brief introduction (Kemper 1999, 43), in which Kemper articulates his three central hypotheses, the first part (pp. 44-48) is devoted to the description and exegesis of Trakl's expressionistic sequential style (Reihenstil). The second part (pp. 48-55) examines the sound-symbolic, motivic and structural repetitions of the poem, while the third part (pp. 55-58) explores the intra- and intertextual references to the rest of Trakl's lyrical oeuvre. These dimensions are brought together concisely in an overall interpretation (pp. 58).

Kemper opens his article with the following hypotheses: Trakl's *Im Winter* belongs (1) "to the early examples of the expressionistic sequential style" and breaks with the characteristics of the classical and romantic tradition of German poetry. It simultaneously realizes (2) "the poetic design of a referenceable winter image of high sensual plasticity", which, however (3) "through its sensual charge and connotative approximation of the motifs" leads these motifs "to lose their everyday linguistic meaning and an autonomization of the poetic texture setting in".

Each of the following three sections is dedicated to the development and support of one of these three hypotheses. The first part explores in detail the realization of the expressionistic sequential style in Trakl's poem and examines its consequences in relation to classical-romantic German poetry. The second part compares the poem with Bruegel's *The Hunters in the Snow* to show that Trakl's poem, like Bruegel's painting, is characterized by "haunting plasticity and suggestiveness". According to Kemper, the sound symbolism – especially assonances and alliterations – as well as motivic and structural repetitions contribute to this effect. In the third part, Kemper shows that "the multiplicity of the image parts and the approximation of the motifs [...] promoted by the form causes a tendency towards the autonomization of the vocabulary that runs counter to its referentiality". This multiplicity is a result of the Trakl-specific connotations established throughout his lyrical oeuvre. In his conclusion, Kemper unites these three lines of argument to assert that Trakl's poem combines the "destruction of traditional poetic meaning and the construction of an autonomous world of signs typical of Trakl" in such a way that their opposition is "'suspended' in sense of a refusal of meaning".

If one attempts an argumentative reconstruction of Kemper's interpretation, one is confronted with a complex argument. Kemper's main thesis – and thus the conclusion of this argument – is the assertion that Trakl's poem ultimately eludes a clear specification of meaning through the interplay of different principles of representation and form.

The justification of this thesis can be reconstructed as a four-part argument (shown in Figure 3–Figure 6 in Appendix 1), with each part consisting of subarguments that have their own intermediate conclusions. The conclusions of the first three main arguments then form the premises of Kemper's central argument. However, the reconstruction necessarily adopts substantive theoretical assumptions – concerning, for instance, the integrability of meaning levels and the criteria of definable meaning – that are not independently argued for in the original text but are nonetheless carried over into the reconstructed argument without being made explicit.

5. Experiments

5.1 Experimental Setup

We conduct experiments with one LLM: Anthropic's Claude-Sonnet-4.5[21]. The model was selected by manually comparing the output of three different LLMs. The key reason for selecting Claude was its ability to account for the structure of the argumentation reconstructions of the interpretations without rigidly adhering to the semantics of individual segments of the prompted examples, unlike other models. We worked with the default temperature of 1, as this yielded the best results in manual, qualitative inspection. All inputs and outputs were generated in German, as we worked with German poems. Additionally, the prompt template already incorporated the modular structure that we consider useful for the continuation of our experiments. For instance, the titles of the poems were entered separately, which allows for future experiments to generate interpretations with or without the inclusion of poem titles. The input to the model was a prompt consisting of a simple task description, an example of an argument-like interpretation and the corresponding poem as well as the poem to be interpreted:

```
You are a literary scholar.
Interpret the following poem in an argumentative form.
- - -
### Orientation:
Below you will find an example that serves only as a structural template,
    not as a content template.
Use the argumentative structure as a guide, but develop new arguments that
     refer exclusively to the new poem.

### Example (structure template only)
Title: {title}
Example: {poem}

Interpretation (example): {interpretation}
- - -
### New Poem
Title: {title_x}
Poem: {poem_x}
```

21. https://www.anthropic.com/news/claude-4-5-sonnet

```
- - -
### Interpretation:
```

During the creation of the prompt templates, we deliberately avoided extensive iterative prompt engineering, as without an algorithmically implemented evaluation procedure and a reliable gold standard data set, prompt optimization becomes prohibitively time-consuming and thus futile (Pichler et al. 2025). As examples (consisting of a poem and its interpretation in the form of an argument-like interpretation), we utilized the three argument-type reconstructions described in section 4, hereafter referred to as reference data. These examples were also employed to iteratively refine the evaluation scheme (cf. subsection 3.3). As test data, we selected six poems – three canonical works and three more contemporary pieces. These are Johann Wolfgang von Goethe's *Über allen Gipfeln*, Hugo von Hofmannsthal's *Manche freilich ...*, Ingeborg Bachmann's *Die gestundete Zeit*, Frederike Mayröcker's *was brauchst du*, Durs Grünbein's *Die leeren Zeichen 19* and Elfriede Gerstl's *balance - balance*. Each of the six poems was interpreted using the template above as a one-shot prompt, with a different reference reconstruction used as an example in each run (in the following marked with 1: Schmidt (1984), 2: Kemper (1999), 3: Meli (1997)). Considering the three argument-like structured interpretations that served as examples, this approach resulted in three interpretations per poem, yielding a total of 18 generated interpretations. These interpretations were first assigned to Strube's levels and then evaluated on the base of the criteria developed in subsection 3.3, using a four-point Likert scale on each interpretation statement by the first and second author of this paper as annotators.[22] To verify the consistency in the application of the criteria, we calculate inter-annotator agreement (Cohen's Kappa; Cohen 1960) as well as the percentage agreement.

Subsequently, we analyse the generated argument-like structured interpretations by examining the agreement with regard to the different levels of argumentation according to Strube – i.e. description, exegesis, interpretation and rule of inference – and the correlation between agreement and the average Likert scores per annotator per level of argumentation.

5.2 Results

Consistency of Evaluation Criteria (Table 1): With regard to the individual argument-like interpretations, we observe an average inter-annotator agreement of 0.74. Average standard deviation values of the Likert scores are 0.61 and 0.64 respectively. Taken together, these scores indicate that the annotators reached a solid and reliable agreement. The moderate standard deviation suggests that while there was some variability in the ratings, the agreement remained within an acceptable range, which supports the robustness of the annotation results – but also shows that the full Likert-range was rarely used. Still, this indicates that the operationalization of Strube's evaluation criteria is reasonably reliable and consistent, given the complexity of the annotated reasoning.

Cohen's Kappa and Average Likert Values per Interpretation and their Ratio (Table 1): The analysis reveals that there is substantial item-level variability. However,

22. Two examples of generated argument-like interpretations can be found in the appendix; see Figure 11 and 11.

Name	IAA	Annotator 1		Annotator 2	
		Mean	Std. Dev.	Mean	Std. Dev.
Grünbein 3	0.9397	3.0800	0.4000	3.0400	0.3512
Goethe 1	0.8622	3.5250	0.6400	3.4359	0.6405
Grünbein 1	0.8385	3.0385	0.4455	3.1538	0.5435
Goethe 3	0.8300	2.5909	0.5032	2.5455	0.5958
Goethe 2	0.8030	2.9574	1.0623	2.9783	1.0644
Hofmannsthal 1	0.7871	3.6970	0.4667	3.6970	0.4667
Hofmannsthal 2	0.7824	3.4130	0.5803	3.4348	0.5012
Mayröcker 2	0.7753	3.2388	0.7404	3.2985	0.7591
Hofmannsthal 3	0.7620	3.8286	0.4528	3.7143	0.5186
Grünbein 2	0.7219	2.8387	0.6323	2.8226	0.6408
Bachmann 1	0.7157	2.8621	0.6394	3.0000	0.7071
Bachmann 2	0.7123	3.2653	0.6701	3.3265	0.6579
Mayröcker 3	0.6883	3.2653	0.5692	3.2653	0.6382
Gerstl 1	0.6779	2.7097	0.7829	2.7419	0.8152
Gerstl 2	0.6647	2.5942	0.8964	2.5507	0.9477
Bachmann 3	0.6385	2.7931	0.4123	3.1951	0.6411
Gerstl 3	0.5842	2.5556	0.6157	2.7222	0.4609
Mayröcker 1	0.5823	2.8750	0.5367	2.8750	0.6124
Average	0.7426	3.0627	0.6137	3.0999	0.6423

Table 1: IAA, Average, and Standard Deviation of Likert scales for both annotators across argument-like structured interpretations.

this variability does not map straightforwardly onto individual authors. For instance, texts by Grünbein span a wide range, from comparatively moderate agreement (0.72 for Grünbein 2) to the highest observed IAA overall (0.94 for Grünbein 3). Mayröcker's texts likewise show marked internal variation, ranging from 0.58 (Mayröcker 1) to 0.78 (Mayröcker 2). Bachmann also exhibits noticeable spread (0.64–0.72), while Gerstl's texts cluster at the lower end of the distribution but still vary substantially (0.58–0.68). By contrast, Hofmannsthal's items show relatively stable and consistently high agreement (0.76–0.79), and Goethe's texts likewise fall within a comparatively narrow and elevated range (0.80–0.86).

The Likert means for both annotators cluster between the middle and the higher end of the scale (approximately 2.5–3.8), and the corresponding standard deviations are relatively homogeneous, mostly between about 0.35 and 1.06. The two annotators show similar dispersion across items, and there is no obvious systematic association between higher IAA and either higher or lower variance in the ratings. Likewise, there is no clear monotonic relationship between IAA and the level of the mean judgments themselves. Overall, the data suggest nuanced, item-specific differences in how interpretable or stable particular argument-like interpretations are, but they do not reveal strong, easily generalizable patterns at the level of individual authors.

Percent Agreement per Argumentation Level (Table 2): To determine the agreement between the annotators with regard to the individual argumentative levels of the generated interpretations, we calculated the percentage agreement and the Pearson correlation efficient of this to the average Likert scores. The results reveal a differentiated pattern across the four argumentative levels. First, average agreement exceeds 75 % in three categories: The highest mean occurs in the rule-of-inference layer (90.89), followed by

Name	Percent Agreement			
	Description	Exegesis	Interpretation	Rule of Inference
Bachmann 1	87.50	46.67	100.00	100.00
Bachmann 2	87.50	68.42	78.57	50.00
Bachmann 3	–	66.67	83.33	–
Goethe 1	88.24	100.00	100.00	100.00
Goethe 2	85.71	70.00	75.00	83.33
Goethe 3	100.00	25.00	75.00	–
Grünbein 1	100.00	66.67	75.00	100.00
Grünbein 2	83.33	76.92	91.67	87.50
Grünbein 3	100.00	94.12	100.00	–
Hofmannsthal 1	0.00	84.00	100.00	100.00
Hofmannsthal 2	100.00	66.67	75.00	87.50
Hofmannsthal 3	100.00	80.00	75.00	100.00
Mayröcker 1	–	68.75	0.00	100.00
Mayröcker 2	87.50	47.06	72.22	85.71
Mayröcker 3	93.75	76.00	66.67	–
Gerstl 1	83.33	66.67	50.00	100.00
Gerstl 2	75.00	58.62	66.67	87.50
Gerstl 3	–	75.00	100.00	–
Average	84.79	68.73	76.90	90.89

Table 2: Percent agreement for each category across poems, sorted alphabetically by poem name, including the arithmetic mean. Empty cells indicate that the LLM did not generate text pertaining to the respective category.

the description layer (84.79 %), and interpretation (76.90 %). Agreement is noticeably lower for exegesis (68.73 %). At the same time, the item-level values exhibit substantial internal variability within all categories. Description agreement, for instance, ranges from as low as 0 % (Hofmannsthal 1) to multiple instances of 100 %. Exegesis spans an interval, from 25 % (Goethe 3) to 94.12 % (Grünbein 3). Interpretation likewise displays considerable dispersion, extending from 0 % (Mayröcker 1) to multiple cases of 100 % (Goethe 1, Grünbein 3, Hofmannsthal 1). Second, missing values (–) appear across two categories, most prominently in the rule-of-inference layer. In five cases, the generated interpretations contain no inferential structures that could be evaluated by both annotators, leading to the absence of a corresponding agreement value. Missing entries in the description category occur only twice and arise exclusively when the model did not produce a descriptive layer at all. Third, the comparatively lower mean agreement in exegesis and interpretation aligns with well-known tendencies in literary scholarship: Such statements are inherently more contestable than descriptive claims. In numerous instances, the model introduced exegetical or interpretative assertions without providing sufficiently clear or consistent descriptive grounding, which resulted in divergent annotator judgments. This mechanism contributes to the broader spread of agreement values in both categories, in contrast to the more structurally constrained description and rule-of-inference layers.

Category-wise Evaluation and Correlations (Table 3): We observe a differentiated pattern in average Likert scores across the annotation layers, with systematically higher evaluations for descriptive components than for the higher argumentative levels: The mean value is highest in the description category (3.48), followed by exegesis (3.15), with lower means in interpretation (3.07) and the rule-of-inference layer (2.66). This

Category	Likert	Correlation of agreement with	
	Mean	A1	A2
Description	3.4768	0.3012	-0.1888
Exegesis	3.1524	0.4480	0.4017
Interpretation	3.0662	0.2589	0.0535
Rule of Inference	2.6600	-0.0458	-0.2061

Table 3: Category-wise evaluation scores. Likert scores are averaged over both annotators, correlation measured between individual Likert scores and percent agreement.

pattern indicates that annotators tended to evaluate the descriptive parts of the interpretations more favorably than the more abstract argumentative components, although the difference between the interpretation and rule-of-inference layers remains small. The low Likert scores for the rule-of-inference layers are related to a problem that already became apparent during the reconstruction process and is also described by Winko et al. (2024): the reconstruction constitutes a highly generalizing supplement in which restrictive conditions are easily overlooked, and whose isolation simultaneously creates the impression of a deductive argumentation. As a result, many of the rules generated by the LLM on this basis are not convincing. In addition, the correlations between the individual Likert scores and the percentage agreement show that a high percentage agreement does not necessarily coincide with high Likert scores. In the description category, correlations with agreement are weakly positive for Annotator 1 (0.30) and moderately negative for Annotator 2 (-0.19), suggesting no stable relationship between agreement and evaluative judgment at this level. The exegesis category exhibits moderate positive correlations for both annotators (Annotator 1: 0.45, Annotator 2: 0.40), pointing to a more consistent alignment between agreement and evaluation than in the description layer. By contrast, the rule-of-inference category displays negative correlations for both annotators (Annotator 1: -0.05, Annotator 2: -0.21), indicating that higher agreement at this level is not associated with higher Likert ratings and may even coincide with more critical evaluations. The interpretation category again shows weak correlations (Annotator 1: 0.26, Annotator 2: 0.05), reinforcing the conclusion that the relationship between agreement and evaluative judgment remains relatively unstable at this global interpretive level.

In aggregate, it can therefore be said: Across categories, percent agreement (PA) is generally high, with the highest values in the rules-of-inference and descriptive layers and the lowest in exegesis, with interpretation showing comparatively intermediate levels of agreement. Likert evaluations show only moderate variation in their mean values across categories, indicating that both annotators applied their judgments in a broadly comparable manner. The correlations between PA and Likert scores differ markedly by argumentative level: They are negative in the descriptive category, clearly positive in exegesis, close to zero in interpretation, and negative again in the rule-of-inference category. Taken together, these results show that the relationship between agreement and evaluative judgments is not uniform across levels but depends on the type of argumentative operation involved.

6. Conclusions

In summary, the workflow developed here for evaluating generated argument-like interpretations of poems appears robust overall, yet its reliability varies across argumentative levels. While the overall moderate to high percent agreement, with notable variation across categories, indicates consistent annotation behavior, the divergent correlations between PA and Likert evaluations show that agreement and evaluative judgments do not align uniformly across descriptive, exegetical, interpretative, and inferential operations. Rather than indicating uniform effects, the correlations suggest category-specific relationships between inter-annotator agreement and perceived quality that admit multiple interpretations. Particularly noteworthy is the positive correlation between PA and Likert scores in the exegetical category, which cautiously suggests that higher annotator agreement may coincide with higher perceived quality of exegetical operations, without implying a strong or universal alignment between agreement and evaluation. In contrast, in the rules-of-inference category, the negative correlation in combination with comparatively high PA values can be read as tentative evidence that instances of agreement, where they occur, tend to coincide with lower Likert scores, potentially suggesting consensual identification of weak or unconvincing rules of inference. Conversely, the near-zero correlation in the interpretative category, alongside a moderate level of agreement, and the mixed correlations observed in the descriptive layer indicate areas where automated interpretations introduce forms of instability that human annotators detect to varying degrees. Together, these findings demonstrate both the sensitivity of the argument-like interpretive framework and the differentiated reliability of automated interpretive outputs depending on the type of argumentative operation involved.

We also see great potential for critical self-reflection of practices of literary studies in the fact that the evaluation of generated interpretation forces researchers to make his/her guiding background assumptions of these practices explicit. A central role in this would probably be played by the analysis of inference rules, as these are the manifestation of framework assumptions that secretly guide interpretation but are rarely made explicit. By explicitly formulating these inference rules in generating interpretations as presented here, much can be learned about the practices of the discipline by evaluating them, without having to carry out the laborious work of partially or fully reconstructing existing interpretations in advance.

7. Future Work

Future work can build upon this study on various levels: For instance, a detailed comparison of existing catalogs of evaluation criteria in terms of their consistency and operationalizability could contribute to refining the approach presented here. Additionally, the number of argumentative levels considered during reconstruction could be gradually expanded to examine the extent to which LLMs can adequately transfer the argumentative structure to new texts at increasing levels of complexity. Furthermore, efforts could be made to improve the skeletal reconstruction of arguments using advanced techniques such as prompt-tuning. By refining prompt design, it may be possible to generate more accurate and nuanced representations of complex literary

interpretations. Additionally, larger test datasets are required to ensure the robustness and generalizability of the findings. These datasets should encompass a broader range of literary-historical epochs and interpretive frameworks, enabling a more comprehensive evaluation of the methodology across diverse contexts. Building on such datasets, future work should also aim to evaluate not only the LLM-generated interpretations, but also the reference interpretations themselves, using the same set of criteria. Comparing these evaluations may yield valuable insights into differences in interpretive practice and argumentative structure between human and machine-generated interpretations. Moreover, if a sufficient amount of manually annotated evaluation data becomes available, a classifier could be trained to automate the evaluation process, thereby enhancing scalability and consistency in future assessments. Finally, it is essential to explore alternative reconstruction approaches that might offer different or complementary perspectives on literary argumentation, contributing to a more versatile and multifaceted framework for the analysis of interpretive texts.

8. Limitations

This study is subject to several limitations. First, the chosen form of reconstruction does not align with common textual practices in literary studies or established literary conventions of representation. As a result, the approach may appear overly simplistic and neglect the specific context of interpretation, such as the intended audience, purpose, or situational relevance. Second, the validity of the study is limited by the fact that the reference data used for comparison was not itself evaluated, leaving potential biases or inaccuracies in the reference data unaddressed. Third, the reproducibility of results poses a challenge due to the non-deterministic nature of large language models and the use commercial models via API. Even with identical inputs and prompts, outputs may vary, making consistent replication difficult. Fourth, the experiment was conducted on a relatively small dataset of 18 examples, limiting the statistical robustness and generalizability of the findings. Fifth, the results cannot be generalized across all LLMs, as the analysis was restricted to a single model, which may not fully represent the capabilities or limitations of other models in the same category. Finally, the study evaluated reconstructed arguments and interpretations rather than directly assessing LLM-generated interpretations. This indirect approach might not capture the full potential or limitations of LLMs in generating literary analyses directly, leaving room for further exploration in future research.

9. Data Availability

Data and code can be found here: https://github.com/AxPic/poem-int-eval. They have been archived and are persistently available at: https://doi.org/10.5281/zenodo.18166524

10. Acknowledgements

We would like to thank the reviewers as well as Janina Jacke and the participants of the General Meeting of the DFG Priority Programme "Computational Literary Studies" in Würzburg for their constructive feedback on earlier versions of this contribution.

11. Author Contributions

Axel Pichler: Conceptualization, Data curation, Investigation, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing

Martin Endres: Validation

Nils Reiter: Formal analysis, Writing – review & editing, dealing with emojis in references

References

Albrecht, Andrea and Lutz Danneberg (2021). "Verstehen, Auslegen, Darstellen und Vermitteln: Literaturwissenschaftliche Interpretationstexte in praxeologischer Perspektive". In: *Doing Interpretation*. Ed. by Johannes Corrodi Katzenstein, Andreas Mauz, and Christiane Tietz. Brill | Schöningh, 23–50. 10.30965/9783657701551_003.

Bamman, David, Kent K. Chang, Li Lucy, and Naitian Zhou (2024). "On Classification with Large Language Models in Cultural Analytics". In: *Proceedings of the Computational Humanities Research Conference 2024*. Ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. https://ceur-ws.org/Vol-3834/paper119.pdf (visited on 12/08/2025).

Beardsley, Monroe C. (1981). *Aesthetics. Problems in the Philosophy of Criticism*. Hackett.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 610–623. 10.1145/3442188.3445922.

Borg, Emma (2025). "LLMs, Turing Tests and Chinese Rooms: the Prospects for Meaning in Large Language Models". In: *Inquiry*, 1–31. 10.1080/0020174x.2024.2446241.

Bowell, Tracy, Robert Cowan, and Gary Kemp (2020). *Critical Thinking: A Concise Guide*. 5th edition. Routledge. 10.4324/9781351243735.

Brandom, Robert (2001). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. 4th edition. Harvard Univ. Press.

Brun, Georg and Gertrude Hirsch Hadorn (2021). *Textanalyse in den Wissenschaften: Inhalte und Argumente analysieren und verstehen*. 4th edition. vdf Hochschulverlag AG an der ETH Zürich. 10.3218/4034-0.

Bühler, Axel (1999). "Die Vielfalt des Interpretierens". In: *Analyse & Kritik* 21 (1), 117–137. 10.1515/auk-1999-0107.

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao (2021). "Evaluation of Text Generation: A Survey". In: *arXiv preprint*. 10.48550/arXiv.2006.14799.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20 (1), 37–46. 10.1177/001316446002000104.

Danneberg, Lutz (1992). "Einleitung. Interpretation und Argumentation: Fragestellungen der Interpretationstheorie". In: *Vom Umgang mit Literatur und Literaturgeschichte*. Ed. by Lutz Danneberg and Friedrich Vollhardt. Metzler, 13–23.

Davies, David and Carl Matheson, eds. (2008). *Contemporary Readings in the Philosophy of Literature: An Analytic Approach*. Broadview Press.

Descher, Stefan (2017). *Relativismus in der Literaturwissenschaft: Studien zu relativistischen Theorien der Interpretation literarischer Texte*. Erich Schmidt Verlag GmbH & Co. KG. 10.37307/b.978-3-503-17461-4.

Descher, Stefan, Jan Borkowski, Felicitas Ferder, and Philipp David Heine (2015). "Probleme der Interpretation von Literatur – Ein Überblick". In: *Literatur interpretieren: Interdisziplinäre Beiträge zur Theorie und Praxis*. Brill | mentis, 11–70. 10.30965/9783957438973_003.

Descher, Stefan and Thomas Petraschka (2018). "Die Explizierung des Impliziten". In: *Scientia Poetica* 22 (1), 180–208. 10.1515/scipo-2018-007.

Döring, Nicola (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. 6th edition. Springer. 10.1007/978-3-662-64762-2.

Hempfer, Klaus W. (2018). *Literaturwissenschaft – Grundlagen einer systematischen Theorie*. Abhandlungen zur Literaturwissenschaft. J.B. Metzler. 10.1007/978-3-476-04700-7.

Herméren, Göran (1983). "Interpretation. Types and Criteria". In: *Grazer Philosophische Studien* 19, 131–161. 10.5840/gps19831923.

Jannidis, Fotis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister (2025). "Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry". In: *Journal of Computational Literary Studies* 4 (1). 10.48694/jcls.4225.

Kemper, Hans-Georg (1999). "Form-(De)-Konstruktion: Poetische Malerei im Reihungsstil". In: *Gedichte von Georg Trakl*. Ed. by Hans-Georg Kemper. Reclam, 43–59.

Koch, Steffen (2025). "Babbling Stochastic Parrots? A Kripkean Argument for Reference in Large Language Models". In: *Philosophy of AI* 1, 19–33. 10.18716/OJS/PHAI/2025.2325.

Köppe, Tilmann (2008). "Konturen einer analytischen Literaturtheorie". In: *Derrida und danach? Literaturtheoretische Diskurse der Gegenwart*. Ed. by Gregor Thuswaldner. VS Verlag für Sozialwissenschaften, 67–83. 10.1007/978-3-531-91822-8_5.

Köppe, Tilmann and Simone Winko (2011). "Zum Vergleich literaturwissenschaftlicher Interpretationen". In: *Hermeneutik des Vergleichs*. Ed. by Andreas Mauz and Hartmut von Sass. Königshausen & Neuman, 305–320. 10.1007/978-3-0348-9261-2_4.

Lin, Chin-Yew (2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81. https://aclanthology.org/W04-1013/ (visited on 12/08/2025).

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig (2023). "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Computing Surveys* 55 (9), 1–35. 10.1145/3560815.

Martus, Steffen (2021). "Interpretieren – Lesen – Schreiben: Zur hermeneutischen Praxis aus literaturwissenschaftlicher Perspektive". In: *Hermeneutik unter Verdacht*. Ed. by

Andreas Kablitz, Christoph Markschies, and Peter Strohschneider. De Gruyter, 45–82. 10.1515/9783110698084-003.

Martus, Steffen and Carlos Spoerhase (2022). *Geistesarbeit. Eine Praxeologie der Geisteswissenschaften*. Suhrkamp. 10.1515/scipo2024-024.

Meli, Marco (1997). "Der Sänger". In: *Gedichte von Gottfried Benn*. Ed. by Harald Steinhagen. Reclam, 87–99.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2001). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics, 311–318. 10.3115/1073083.1073135.

Petraschka, Thomas and Stefan Descher (2019). *Argumentieren in der Literaturwissenschaft. Eine Einführung*. Reclam Verlag.

Pichler, Axel, Janis Pagel, and Nils Reiter (2025). "Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies". In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*. Ed. by Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni, and Janis Pagel. Association for Computational Linguistics, 32–46. 10.18653/v1/2025.latechclfl-1.5.

Schmidt, Jochen (1984). ""Sobria ebrietas". Hölderlins "Hälfte des Lebens"". In: *Gedichte und Interpretationen. Band 3: Klassik und Romantik*. Ed. by Wulf Segebrecht. Reclam, 256–267.

Sellars, Wilfrid (1953). "Inference and Meaning". In: *Mind* 62.247, 313–338. ISSN: 00264423, 14602113. http://www.jstor.org/stable/2251271 (visited on 12/19/2025).

Strube, Werner (1992). "Über Kriterien der Beurteilung von Textinterpretationen". In: *Vom Umgang mit Literatur und Literaturgeschichte*. Ed. by Lutz Danneberg, Friedrich Vollhart, Hartmut Böhme, and Jörg Schönert. Metzler, 185–210. 10.1007/978-3-476-03386-4_8.

Susteck, Sebastian and Christoph Perder (2023). "Schreiben durch Künstliche Intelligenz. ChatGPT und automatisierte Lyrikanalysen". In: *MiDU - Medien im Deutschunterricht*, 1–20. 10.18716/OJS/MIDU/2023.0.2.

Toulmin, Stephen E. (2003). *The Uses of Argument*. 2nd edition. Cambridge University Press.

Winko, Simone (2015). "Zur Plausibilität als Beurteilungskriterium literaturwissenschaftlicher Interpretationen". In: *Theorien, Methoden und Praktiken des Interpretierens*. Ed. by Andrea Albrecht, Lutz Danneberg, Olav Krämer, and Carlos Spoerhase. De Gruyter, 483–512. 10.1515/9783110353983.483.

Winko, Simone, Stefan Descher, Urania Milevski, Merten Kröncke, Fabian Finkendey, Loreen Dalski, and Julia Wagner (2024). *Praktiken des Plausibilisierens: Untersuchungen zum Argumentieren in literaturwissenschaftlichen Interpretationstexten*. Göttingen University Press. 10.17875/gup2024-2639.

Zabka, Thomas (2008). "Interpretationsverhältnisse entfalten. Vorschläge zur Analyse und Kritik literaturwissenschaftlicher Bedeutungszuweisungen". In: *Journal of Literary Theory* 2 (1), 51–69. doi:10.1515/JLT.2008.005.

Appendix 1: Reconstruction of the Core Argumentation of *Kemper*

- **Premise 1** (= P_1): The poem consists of three four-line stanzas with different individual images from the natural and human world.
- **Premise 2** (= P_2): With the exception of three enjambments, the end of the sentence and the end of the verse coincide in the poem, which reinforces the pauses between the images.
- **Premise 3** (= P_3): The three enjambments only connect main clauses and do not break up sentences.
- **Premise 4** (= P_4): The simple, uniform sentence structure supports the clear separation of the images.
- **Intermediate Conclusion 1** (= C_1): The poem realizes a new poetic image in each verse.
- **Material rule of inference** (= R_1): If images in a poem are arranged as a strict sequence of independent units without syntactic dependence, this constitutes a sequence of independent individual images, which is characteristic of the expressionist sequential style.
- **Intermediate Conclusion 2** (= C_2): The poem realizes the expressionistic sequential style.
- **Premise 5** (= P_5): The expressionist sequence style suspends the level of symbolic meaning.
- **Conclusion** (= Final C): The poem has no symbolic meaning.

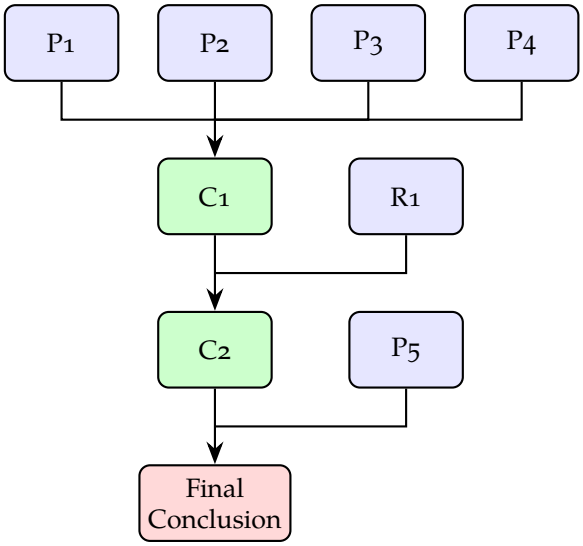

Figure 3: Reconstruction of Argument 1.

- **Premise 1** (= P_1): The first and second verses of the poem show a repetitive but varied optical sequence of movement from the ground to the sky and from the sky to the ground and back.
- **Premise 2** (= P_2): The first verse of the poem introduces two antonymic assonance groups ('a' and 'ei'), which lead to a blending of optical and haptic perceptions.
- **Premise 3** (= P_3) The contrasting and at the same time analogous perceptual values thus created are continued and intensified denotatively and connotatively in the following verses of the poem.
- **Intermediate Conclusion 1** (= C_1): The poem is characterized by sound-symbolic, motivic and structural repetitions and correspondences.
- **Material rule of inference** (= R_1): If a poem exhibits a high degree of sound-symbolic, motivic and structural repetitions and correspondences, then it potentially creates an impression of iconicity analogous to painting.
- **Intermediate Conclusion 2** (= C_2): The poem creates an impression of iconicity analogous to painting.
- **Material rule of inference** (= R_2): If a text achieves a painting-like iconicity and its images can be related to real winter scenes, then it offers a poetic winter image of high sensual plasticity that can be referentialized.
- **Premise 4:** = P_4) The images in the poem can be related to real winter scenes.
- **Conclusion** (= Final C): The poem offers a poetic winter image of high sensual plasticity that can be referentialized.

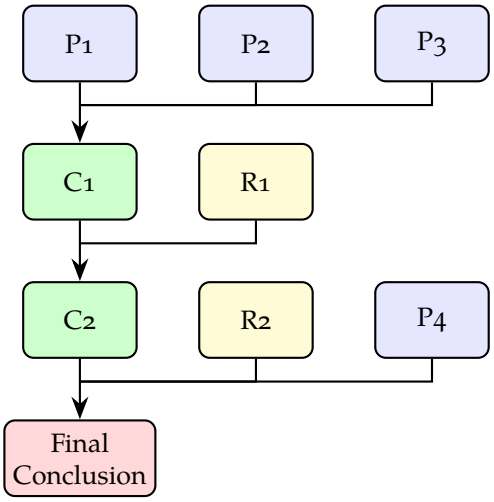

Figure 4: Reconstruction of Argument 2.

- **Premise 1** (= P_1): Trakl uses a limited vocabulary of images and motifs, which he combines and varies in different poems.
- **Premise 2** (= P_2): The recurring use of certain images and motifs in different poems creates a Trakl-specific intertextuality.
- **Premise 3** (= P_3): The poem contains images and motifs that also recur in other poems by Trakl.
- **Intermediate Conclusion** (= C_1):The poem participates in the Trakl-specific intertextuality.
- **Premise 4** (= P_4): In the poem, numerous inter- and intratextual relations between the image parts and an approximation of the motifs are present.
- **Material rule of inference** (= R_1): If, in a poem, numerous inter- and intratextual relations between the image parts and an approximation of the motifs are present, this tends to lead to an autonomization of its vocabulary.
- **Conclusion** (= Final C): The poem has a tendency toward an autonomization of its vocabulary.

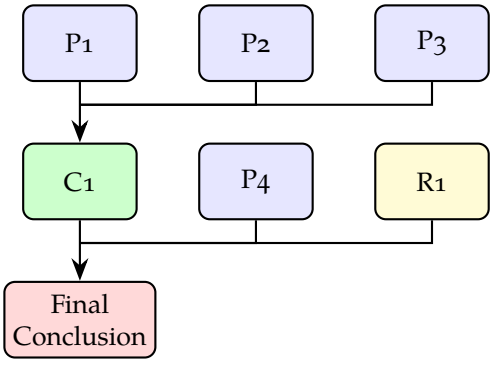

Figure 5: Reconstruction of Argument 3.

- **Conclusion of argument 1** (= C_1): The poem has no symbolic meaning.
- **Conclusion of argument 2** (= C_2): The poem offers a poetic winter image of high sensual plasticity that can be referentialized.
- **Conclusion of argument 3** (= C_3): The poem has a tendency toward an autonomization of its vocabulary.
- **Premise 4** (= P_4): These three aspects constitute different, non-integrable semantic levels of the poem.
- **Intermediate Conclusion** (= C_4): The poem offers different meanings on different representational levels.
- **Material rule of inference** (= R_1): If a poem offers different meanings on different representational levels, then it eludes a clear definition of its meaning.
- **Conclusion:** (= Final C): The poem *Im Winter* eludes a clear definition of its meaning.

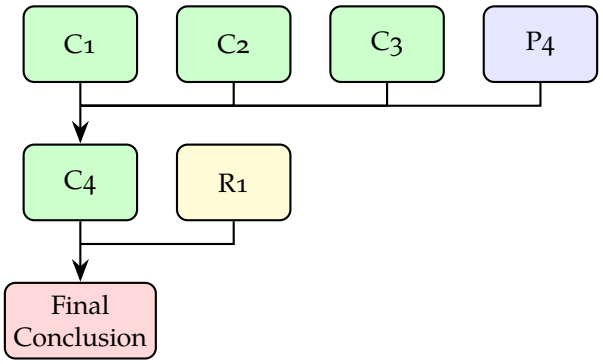

Figure 6: Reconstruction of Argument 4.

Appendix 2: Generated Argument-Like Interpretation of Hofmannsthal's *Manche freilich ...*, prompted with Meli's Reconstruction, and Gerstl's *balance–balance*, also prompted with Meli's Reconstruction

The following sections show two LLM-generated argument cores that have been manually evaluated in this paper. We show the interpretation core with the highest agreement and the highest average Likert scores (Hofmannsthal 3), as well as the interpretation core with a high agreement and correspondingly lower Likert scores (Gerstl 3).

Hofmannsthal 3

```
1  Die ersten beiden Strophen des Gedichts »Manche freilich ...« stellen zwei
     gegensätzliche soziale Existenzformen einander gegenüber.
2  Die eine Gruppe muss „drunten sterben" (V. 1) und liegt „mit schweren
     Gliedern / Bei den Wurzeln des verworrenen Lebens" (V. 5-6), während die
     andere „droben" (V. 3) wohnt und bei „Sibyllen" und „Königinnen" (V. 8)
     sitzt.
3  Die räumliche Opposition von „drunten" und „droben" korrespondiert mit der
     Gegenüberstellung von „schwer" und „leicht".
4  Die Wiederholung des Wortes „manche" bzw. „andre" etabliert eine
     dualistische Weltordnung.
5  Diese dualistische Weltordnung wird in der dritten Strophe problematisiert
     .
6
7  Ein „Schatten" (V. 11) verbindet die getrennten Sphären miteinander, und
     „die leichten sind an die schweren / Wie an Luft und Erde gebunden" (V.
     13-14).
8  Die Naturmetapher „Luft und Erde" suggeriert, dass diese Verbindung
     existenziell und unauflösbar ist.
9  In der vierten Strophe wechselt das Gedicht von der allgemeinen
     Beobachtung zur subjektiven Perspektive eines lyrischen Ich.
10 Das lyrische Ich kann „ganz vergessener Völker Müdigkeiten" nicht „abtun"
     (V. 15-16) und vermag das „stumme Niederfallen ferner Sterne" nicht von
     seiner „erschrockenen Seele" fernzuhalten (V. 17-18).
11 Das lyrische Ich artikuliert eine existenzielle Verbundenheit mit dem
     Leiden anderer und mit kosmischen Vorgängen.
12 Die Unmöglichkeit der Abgrenzung deutet auf eine ethische Verantwortung
     hin, die das privilegierte Individuum nicht ablegen kann.
13
14 Die Schlußstrophe universalisiert diese Erkenntnis durch die Feststellung,
      dass „viele Geschicke" nebeneinander existieren (V. 19).
15 Das „Dasein" verwebt alle Schicksale „durcheinander" (V. 20), wodurch
     individuelle Existenz als Teil eines überindividuellen Zusammenhangs
     erscheint.
```

```
Das lyrische Ich behauptet, sein „Teil" sei „mehr als dieses Lebens /
  Schlanke Flamme oder schmale Leier" (V. 21-22).
Die Metaphern „Flamme" und „Leier" verweisen auf ästhetische Autonomie und
   individuelles Leben.
Die Negation dieser Metaphern bedeutet eine Absage an die Vorstellung
  einer in sich geschlossenen ästhetischen oder individuellen Existenz.
Das Gedicht »Manche freilich ...« verhandelt das Verhältnis von
  privilegierter und unterprivilegierter Existenz sowie die Verantwortung
  des Individuums gegenüber dem kollektiven Leiden.
Wenn ein Gedicht soziale Gegensätze thematisiert und zugleich die Rolle
  des Dichters reflektiert, kann es als sozialethischer und poetologischer
   Text gelesen werden.
Das Gedicht »Manche freilich ...« ist ein sozialethischer und
  poetologischer Text.

Das
   lyrische Ich gehört offenbar zur privilegierten Gruppe der „Leichten".
Das lyrische Ich kann sich jedoch nicht von der Verbindung zu den
  „Schweren" lösen.
Diese unlösbare Verbindung wird nicht als Begrenzung, sondern als
  Erweiterung der eigenen Existenz dargestellt.

Das Gedicht präsentiert eine Poetik der Empathie und der kollektiven
  Verantwortung, die über ästhetischen Autonomieanspruch hinausgeht.
Das Gedicht »Manche freilich ...« steht damit in der Tradition einer
  ethisch engagierten Dichtung, die soziale Gegensätze nicht ausblendet,
  sondern in die poetische Selbstreflexion integriert.
```

Gerstl 3

```
Die erste Strophe des Gedichts »balance - balance« verneint explizit, dass
   räumliche Distanz eine »lösung« für unbenannte Probleme darstellt (V.
  1-4).
Die zweite Strophe wiederholt und erweitert das Motiv der räumlichen
  Entfernung, indem sie das »sitzen auf stufen weit weg« als Zustand der
  Schwebe charakterisiert (V. 5-9).

Das Gedicht »balance - balance« verhandelt das Verhältnis von physischer
  Ortsverlagerung und innerer Problembewältigung.
Die physische Distanz (»rom« vs. »wien«) wird als unwirksame
  Fluchtbewegung entlarvt, da die Probleme »in meinem kopf« verbleiben.
Das zentrale Motiv der »schwebe« (V. 8) korrespondiert mit dem Titel
  »balance - balance« und bezeichnet einen prekären Gleichgewichtszustand.
Dieser Schwebezustand wird durch zwei gegensätzliche Bewegungen
  konstituiert: die triviale Ablenkung durch »lästige papier-lüsterne
  ameisen« (V. 7) und die bedrohlichen »abstürze« (V. 9).

```

9 Das Gedicht arbeitet mit einer Gegenüberstellung von scheinbaren und tatsächlichen Abgründen: Die Flugzeuge schweben »über den so genannten wirklichen abgründen« (V. 11), während das lyrische Ich über »gar nicht unwahrscheinlichen abstürzen« schwebt (V. 9).
10 Die Qualifizierung der Abgründe als »so genannte wirkliche« (V. 11) stellt die Hierarchie zwischen äußerer und innerer Bedrohung infrage.
11 Die Flucht erfolgt zweifach: räumlich von Wien nach Rom und mental durch »wegschauen« (V. 12) auf »kleine käfrige buchstaben« (V. 13).
12 Das Schreiben selbst wird als Ablenkungsstrategie kenntlich gemacht, die Teil des prekären Gleichgewichts ist.
13 Das Gedicht »balance - balance« präsentiert einen Zustand permanenter Instabilität, in dem weder räumliche Flucht noch künstlerische Ablenkung eine genuine Lösung bieten.
14 Die konsequente Kleinschreibung und die parataktische Syntax verstärken formal den Eindruck der Desorientierung und fehlenden Hierarchie.
15 Die Wiederholung des Titelworts »balance« deutet auf die Notwendigkeit ständiger Ausbalancierung hin, nicht auf einen erreichten Gleichgewichtszustand.
16 Das Gedicht »balance - balance« verweigert jede Form von Transzendenz oder Auflösung und beharrt stattdessen auf der Unausweichlichkeit des prekären Schwebezustands als existenzielle Grundsituation der Moderne.