Universiteit Leiden
The Netherlands

# Bachelor Computer Science & Datascience and Artificial Intelligence

A Cross-Cultural Analysis of Humor Detection and Interpretation in Large Language Models

Kai Epprecht
[S3632369]

Supervisors:
Dr. M.J. van Duijn & Lennard Froma

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/11/2025

**Abstract**

Humor is a cross-cultural yet culturally specific mode of human communication that reveals much about the structure of language, common social values, and unspoken cultural beliefs. Recent developments in Large Language Models (LLMs) have shown impressive text-generation and text-comprehension abilities, yet the capacity to understand and assess humor remains unclear. In this thesis, we explore how humor is interpreted and evaluated across various cultural and linguistic environments using LLMs.

The thesis builds on previous research, presenting a cross-linguistic and cross-cultural comparison of humor detection in LLMs. Although both English and German are Western cultures, they differ significantly in linguistic structure and humor conventions. The humor of Japan adds another non-Western cultural dimension to the comparison between the two linguistic and cultural borders. Two main datasets were built: one consisting of German humor collected from various sites, such as *witze.tv* or *hahaha.de*, and another Japanese humor corpus, also gathered from multiple sites, as well as the Dajare Database provided by Araki Labs. A non-humorous control corpus was added to each dataset, which allowed quantitative and qualitative assessment of the models' performance.

Multi-tiered evaluation frameworks established using prior research assessed the ability of pre-trained LLMs to identify humorous content (Tier 1), evaluate funniness and other dimensions related to it (Tier 2), and to provide explanations that reflect an understanding of cultural or contextual influences (Tier 3) of their humor judgments. Following the unexpected demonstration of high levels of performance by these LLMs during the first round of tasks to classify humor, a multi-tiered follow-up approach was added to the original evaluation framework. Tier 4 tests for performance at detecting structural similarities between control responses and humorous responses. Tier 4 includes controls that are structured similarly to humorous responses in terms of dialogue format, punctuation and length. Tier 5 employs semantically neutralized counterfactual versions of humorous responses to determine if LLMs responded to the meaning of the humor or simply to the structure of the response. The addition of this additional tier allows the thesis to test both cross-cultural humor recognition as well as whether apparent performance is due to actual semantic and cultural understanding.

The study can be applied to the interpretation of computational humor and offers the methodological basis of creating culturally adaptive language models. The results can be used to determine the fairness and inclusivity of AI systems involved in cross-cultural communication across multiple languages.

# Contents

# 1 Introduction

## 1.1 Background and Motivation

Humor is a deeply human and culturally ingrained mode of communication that not only reveals the linguistic frameworks but also standard social norms, expectations, and values. Computational explanations of humor have been a classic problem of Natural Language Processing (NLP) and Artificial Intelligence (AI). Large Language Models (LLMs), such as GPT and other transformer-based models, have demonstrated high linguistic fluency. However, their ability to understand, generate, and analyze humor remains unclear. Understanding humor requires a receptivity to context, implicit meaning, and culture, all of which go beyond the surface-level perception of the text.

There have been significant advancements in computational humor detection, particularly in the ability to identify humor across diverse datasets, including Reddit, Twitter, and extensive joke collections. However, nearly all research in this area focuses on English-language humor. As a result, evaluation setups often implicitly assume some degree of cultural homogeneity and universal humor characteristics [MSP10, MS06, YLDH15], which severely limits the ability to assess how humor is interpreted across cultures and languages, particularly with regards to humor from non-western cultures, such as Japanese humor, whose use of cultural references, linguistic puns, and polite expressions are based upon fundamentally different principles than those found in Western cultures [KS09, Tsu11].

A further complicating factor is that strong model performance on humor detection tasks does not always demonstrate an actual ability to understand humor. Models may identify and utilize obvious patterns in how a particular dataset was formatted. As such, exploiting superficial regularities strongly improves model accuracy. These artifacts include: dialogue formatting; punctuation use; structural setup for establishing a stereotype; and the length of sentences in either humorously written passages versus non-humorously written passages. As a result, it appears that these models have demonstrated an "understanding" of humor based solely upon their ability to recognize common surface forms, as opposed to their ability to recognize the underlying mechanisms of what make something funny.

## 1.2 Problem Statement

Although the detection of humor in English has been widely studied, cross-cultural humor comprehension has received the least attention. LLMs, which have been trained on data mostly centered around English, might fail to properly decode humor in other languages, particularly humor based on wordplay, implicit meaning, or sociocultural allusions. The question, therefore, emerges: *to what extent do LLMs generalize their humor recognition capabilities across cultures and languages?* In addition to the cross-cultural differences in humor, there are also several methodological concerns when testing for humor in LLMs. If, as a result of the way they were written (style or structure), the humorous and non humorous texts have systematic differences, then models may achieve high classification accuracy simply due to structural shortcuts, rather than a true understanding of humor in a semantically or culturally based way. The initial stages of the current study suggested that models might be using structural shortcuts instead of semantic or cultural-based humor understanding. These possibilities motivated the development of additional adversarial test controls

for a later extension of the basic evaluation framework.
This thesis builds on prior work evaluating LLMs' humor detection and explanation capabilities—most notably the prompt-based probing of ChatGPT's joke generation, explanation, and classification behavior by Jentzsch and Kersting [JK23], and the two-tier evaluation framework for humor detection and interpretation proposed in Álvarez Jorge's thesis [AJ24]. It explores the performance of LLMs on multilingual and culturally specific joke corpora. It particularly compares variations in the comprehension of humor between Western (German) and non-Western (Japanese) humor, discussing their linguistic play, offensiveness, and cultural awareness. The study examines whether the apparent humor competence of state-of-the-art LLMs reflects robust cross-cultural understanding, or whether it is partly driven by reliance on surface-level textual cues.

## 1.3 Scientific and Societal Relevance

This research contributes to two domains. Scientifically, it enhances our understanding of cross-linguistic humor modeling and highlights potential biases in large-scale language models trained on culturally imbalanced data. Societally, this question becomes increasingly relevant as AI systems are more frequently deployed in cross-cultural and multilingual interactions. For instance, in translation, dialogue systems, and creative media, ensuring that they interpret humor appropriately and respectfully becomes crucial. Misinterpretation of humor can lead to cultural misunderstandings, reinforce stereotypes, or cause unintended offense. The study provides scientifically based research into the evaluation of humor and the methodology of humor evaluation. The study examines if high benchmark scores are a true indication of a model's ability to understand humor. The examination of this question is especially important when developing culturally sensitive tasks because structurally biased evaluation sets can lead to an inflated assessment of model competence. A better designed benchmark design will help differentiate between a model's ability to provide a true semantic/cultural interpretation of a joke and a model providing a shortcut-based classification.

## 1.4 Thesis Overview

This thesis is structured as follows. Section 2 introduces the key theoretical concepts relevant to humor, cultural linguistics, and LLMs. Section 3 presents a literature review of computational humor detection and cross-cultural NLP. Section 4 states the research question, as well as the hypotheses formed based on findings discussed in the previous section. Section 5 details the data collection, preprocessing, and experimental methodology, including the construction of German and Japanese humor datasets and the evaluation pipeline. Section 6 presents and discusses the findings of the experiment through a multi-tiered approach outlined in Section 5, interpreting results at each tier and drawing cross-tier comparisons. Finally, sections 7, and 8 summarizes the findings, discusses limitations, and proposes directions for future research on cross-cultural humor analysis in AI.

This bachelor thesis is conducted at Leiden University's LIACS department under the supervision of Dr. Max van Duijn and Lennard Froma.

# 2 Definitions and Theoretical Background

This section outlines the theoretical foundations necessary to interpret the empirical findings of this thesis. Rather than providing an exhaustive overview of humor theory or language modeling, the discussion focuses on those concepts that are directly relevant to *cross-cultural humor interpretation by LLMs*. The presentation builds upon established definitions used in prior work, in particular the evaluation framework introduced by Álvarez Jorge [AJ24], and extends them to a multilingual and cross-cultural setting.

## 2.1 Humor as a Linguistic and Cultural Phenomenon

Humor is commonly described as a cognitive response to incongruity, ambiguity, or norm violation. Classical humor theories such as *Incongruity Theory*, *Superiority Theory*, and *Relief Theory* offer high-level accounts of why humor occurs [Att94, Mor09]. Although the above theoretical approaches serve as an important conceptual grounding, they remain too abstract to fully explain how humor manifests in various languages and cultures. As Álvarez Jorge [AJ24] comments, the ability to detect humor in actuality goes well beyond identifying incongruent aspects; it also depends on culturally specific scripts, social hierarchy, and audience expectations that vary widely across communities. Therefore, in this thesis, humor will be considered a *culturally situated linguistic practice* [MS06, Apt85]. To understand humor, there needs to be shared background information, implied social norms, and some level of familiarity with the language-specific mechanisms, such as word play, irony, and pragmatic implicature [Att94, Mor09]. Álvarez Jorge [AJ24], in addition to noting that even within a single language, detecting humor can be obscured by small nuances that require localized cultural context, states that this problem is greatly heightened when dealing with both cross-cultural and cross-linguistic contexts. These properties make humor a particularly demanding test case for computational language models, since these models need to move beyond simply processing what is on the surface of texts to capturing culturally embedded meanings [JK23].

## 2.2 Cross-Cultural Dimensions of Humor

Humor perception varies substantially across cultures. What is considered funny in one cultural context may be confusing, unremarkable, or even offensive in another, as humor relies on culturally specific scripts, taboos, and communicative conventions [Dav90, Apt85]. German and Japanese humor illustrate this contrast particularly clearly. German humor often emphasizes irony, understatement, and logical or semantic incongruity, whereas Japanese humor frequently relies on phonological wordplay, such as *dajare*, and on socially structured interaction patterns, including role inversion and politeness violations [How04, KS09, Tsu11, Hal21]. These differences are not merely stylistic but reflect deeper contrasts in linguistic structure and pragmatic norms. From a computational perspective, cross-cultural humor presents a major challenge. Humor mechanisms based on phonology or culturally specific assumptions often do not survive literal translation, and models trained on dominant Western data may misinterpret or overlook such humor entirely. A further theoretical framework for examining the nature of meaning comes from the field of intercultural communication. This interpretive lens describes meaning as varying in the degree to which it is carried by explicit verbal coding versus shared context. In this sense, Japanese communication has often been discussed as being more contextually dependent than prototypically

low-context Western settings [Hal76]. However, these types of national categorizations should be approached with caution, as later studies have questioned their empirical consistency [KRM11]. Japanese humor has also been described as being more interactionally situated than many of the ready-made joke traditions seen in some Western contexts. Japanese joking has been characterized as more closely tied to interpersonal familiarity, low-formality situations, and conversational context, rather than centering on jokes told to unknown audiences, [NR17]. This emphasizes the point that cross-cultural differences in humor are both linguistic, social, and pragmatic. Evaluating LLMs across culturally distinct humor traditions therefore provides insight into the extent to which their humor recognition abilities generalize beyond their training distributions.

### 2.3 Conceptual Challenges in Cross-Cultural Humor Detection

Computational humor detection often frames humor through a binary classification task: humorous (1) or non-humorous (0). Earlier studies approached humor detection through manually designed features such as lexical incongruity, sentiment shifts, or syntactic structure [MS06]. However, recent deep learning models implement contextual embeddings and transformer architectures in order to detect subtler humor cues [WS19]. Although these methods achieve strong results on standard benchmark datasets, they are typically reliant on monolingual, culturally homogeneous corpora. Consequently, systems may learn surface-level indicators of humor, but are unable to capture the underlying cultural humorous intent. Such issues may especially stand out when evaluating across multilingual settings, given that humorous content often varies by language. Another issue related to a theoretical perspective on humor detection benchmarks for computational systems are that benchmarks may combine humor with various other attributes of language. Since humorous texts often have different structural characteristics than control texts (e.g., greater use of exclamation points, shorter lengths), if a model learns to identify those structural differences as indicators of whether a particular piece of text is humorous, then its ability to correctly categorize texts would not necessarily indicate that it understands what makes something funny.

### 2.4 Large Language Models and Cultural Knowledge

A Large Language Model (LLM) is a subcategory of AI, designed to process and generate human language. LLMs are trained on vast text collections using transformer architectures, allowing them to capture statistical and linguistic patterns and contextual relationships [BMR+20]. While LLMs can often identify humor from surface-level textual patterns, their performance is weaker when interpretation depends on culturally specific context, norms, or underrepresented linguistic structures [JK23, N+25, PPJ+25]. Earlier studies indicate system outputs can carry social and value-laden distortions passed down from training inputs [SCNP21]. Testing comedic understanding across multilingual samples thus provides a valuable lens through which to study how such cultural knowledge is encoded within large-scale language models.

## 3 Related Work

This section reviews prior research relevant to the present study, with a focus on computational humor detection, multilingual and cross-cultural humor analysis, and the evaluation of humor

capabilities in LLMs. Rather than providing a comprehensive overview of NLP research, the discussion highlights limitations in existing work that motivate the cross-cultural focus of this thesis.

## 3.1 Computational Humor Detection

Whereas Section 2.3 introduced humor detection as a conceptual and cultural challenge, this section reviews concrete computational approaches and empirical results reported in the literature. As discussed in Section 2, early computational approaches to humor focused on surface-level cues such as lexical incongruity, sentiment polarity, and syntactic structure. In practice, these ideas were implemented through feature-based classifiers applied to short texts such as one-liners, tweets, and joke corpora [MS06, YLDH15]. These methods demonstrated that humorous text often exhibits higher lexical diversity and unexpected word associations, enabling reliable classification in constrained settings. The introduction of deep learning architectures marked a significant advance. Transformer-based models such as BERT and RoBERTa improved context sensitivity and semantic coherence, achieving strong performance on humor benchmarks derived from social media, joke repositories, and multimodal speech datasets [WS19, HRZ+19]. However, these datasets are overwhelmingly English-language and culturally homogeneous, and the incorporation of acoustic or visual cues, datasets typically implement humor as a binary (humorous vs. non-humorous) or event-based signal. As a result, they rarely model audience-dependent variation in humor perception, which can depend on factors such as context, social norms, individual background, and, importantly for this thesis, socio-cultural and linguistic conventions

## 3.2 Limitations of Multilingual Humor Studies for Cross-Cultural Evaluation

While recent studies have examined humor detection beyond English, few have compared across cultures. Current studies typically implement multilingual humor detection by extending monolingual classification pipelines. For instance, Russian [BBBB19] has been analyzed independently; however, humor is treated as a language-agnostic label. Multilingual studies often rely on translation, shared embeddings, or language-specific adjustments [WS19], rather than on modeling local norms, speaking patterns, or audience preferences. As a result, linguistic traits blur with cultural context, leaving uncertainty about whether models truly grasp humor across cultures.

Whilst research on multilingual datasets exists, humor detection is usually implemented as a basic binary task or a constrained editing problem, as in creative headline modification benchmarks such as Humicroedit [HKG19]. While these datasets represent essential advances in task formulation and scale, they do not capture models' aptitude for grasping the culturally grounded intent of humor, a limitation emphasized by audience-centered work such as ManzaiSet [KNR25]. Consequently, there is a lack of a commonly adopted framework for evaluating humor across culturally distinct humor traditions. This limitation motivates the need for an evaluation framework that ensures models are challenged on humor with different cultural contexts, testing models based on their reasoning, rather than a binary classification.

### 3.3 Culturally Specific Humor and Audience-Centered Datasets

Although many current computational humor benchmarks evaluate humor through a binary classification system, recent studies have begun shifting towards methods that consider cultural context and audience perception. One key case is *ManzaiSet*, developed by Kawamura et al. [KNR25], it represents the first significant multimodal dataset, combining video and audio responses to Japanese *manzai* comedy shows. Rather than assessing textual humor, ManzaiSet captures synchronized facial movements and vocal changes from 241 Japanese participants during comedy performances. From an implementation perspective, ManzaiSet is notable for its audience-centered approach; humor is measured by observable behavior rather than textual labels. This allows for a detailed study of viewers' responses to jokes rooted in a unique comedy form, characterized by structured *boke–tsukkomi* exchanges, dialect wordplay, and corrective interaction patterns. Such elements help define shared expectations about humor timing and interpretation, a limitation that textual transcripts face. While ManzaiSet does not test computational humor models or cross-cultural performance directly, it presents a shift away from text-only humor benchmarks. By showing how culture-specific traits can be encoded into data design and evaluation methods, it offers new insights beyond studies focused solely on model-centered humor evaluation.

### 3.4 Large Language Models and Humor Understanding

LLMs such as GPT, PaLM, or Llama have demonstrated a strong ability to detect, clarify, and generate jokes, not as a result of their primary objective, but as a consequence of being trained on large-scale language modeling and instruction tuning [BMR+20]. Such capabilities have sparked growing interest in using humor analysis as a means to benchmark advanced language competency. However, new evidence suggests that these humor-related results may mask significant shortcomings.

Several studies have suggested that LLMs rely on surface-level pattern textual cues, rather than grasping the humor's pragmatic mechanisms. For instance, Jentzsch and Kersting [JK23] show that, while ChatGPT has the ability to generate joke-style responses, along with reasonable justifications, it falls short on the ability to distinguish between genuine humorous content and non-humorous texts that are structurally similar. These findings suggest that evaluation frameworks must move beyond standard humorous-versus-neutral classification and test whether models continue to identify humor when non-humorous controls are made structurally or semantically similar to jokes. Further research conducted by Valois et al. [RVF25] introduces a quantitative analysis framework for humor detection in stand-up routines; their results reveal that leading models still achieve modest accuracy, highlighting persistent challenges in implementing an underlying understanding of humor in computational systems.

Benchmark-based evaluations further indicate that LLMs' performance in humor detection is strongly influenced by their general-purpose reasoning ability rather than by their culturally grounded awareness. Such findings can be seen in HumorBench [N+25], a benchmark developed to assess models' qualitative capabilities for explaining sophisticated cartoon humor. Results show that model performance aligns closely with established STEM-oriented reasoning benchmarks, such as ARC-AGI and GPQA, indicating a strong transfer of general analytical reasoning skills to humor explanation tasks. At the same time, models fail on many examples that require subtle implicit conceptual leads, obscure references, or missing background knowledge to explain why a joke is humorous, highlighting a gap between strong formal reasoning performance and robust humor

comprehension. Similarly, the Oogiri study further highlights the dependence of humor-related tasks on dataset construction. In their work on the Japanese creative-response humor task, they argue that the ability of LLMs to understand humor goes beyond simple pattern recognition; instead, the understanding of humor by LLMs is based upon both cultural and context-based nuance. As important as their arguments regarding the contextual and cultural nuances involved in humor understanding are, however, the authors of the Oogiri study also identify methodological shortcomings in previous humor collections, namely, that many of these collections were developed with fairness concerns and structural biases, and therefore while the systems being studied performed well in terms of task completion, the evaluations used to assess them did so in settings that made it possible to create artificial distinctions between humorous and non-humorous responses that may not accurately represent true humor comprehension. While the Oogiri study is limited to humor within a Japanese single culture (and limits themselves to evaluating the funniness of responses using the *Oogiri* dataset), it does support the necessity to develop humor benchmarks that account for potential artifact and shortcuts in the data when attempting to determine the extent to which LLMs comprehend humor [MKTO25].

Research into non-English humor detection reveals that results are strongly influenced by dataset size, construction, and cultural alignment. Blinov et al. (2019) found that models achieved strong performance gains with large, lexically controlled datasets, suggesting that prior shortcomings in non-English humor detection stem from the lack of high-quality annotated humor datasets rather than modeling limitations [BBBB19]. While Blinov et al. (2019) did not conduct a cross-lingual evaluation, their insight has driven further research such as ManzaiSet [KNR25], which illustrates the association between humor perception and culture-specific norms and audience expectations. These studies imply that systems trained predominantly on data from a single linguistic or cultural context — whether that is English or otherwise — may struggle to generalize humor understanding across cultural contexts. The broader issue is not simply a matter of data quality or volume, but of cultural transferability: training on resources grounded in one language or cultural tradition does not automatically confer the ability to interpret humor rooted in another. This is particularly consequential for the present study, as the models evaluated here are largely trained on English-dominated corpora, yet are assessed on German and Japanese humor traditions that draw on distinct linguistic structures, pragmatic conventions, and cultural assumptions. Such limits further grow due to social and normative biases ingrained in training material, skewing the model's interpretation of humor tied to specific communities [SCNP21].

These results suggest that current tests of LLMs' grasp of humor still struggle to assess how well they generalize across cultures. The multi-tier humor evaluation framework introduced by Álvarez Jorge [AJ24] presents an essential shift from binary classification judgments to explanation-based analysis. The present thesis aims to expand upon this approach by applying it directly in a cross-cultural context, comparing responses to German and Japanese jokes, not just highlighting differences but revealing where models fail culturally. As such, it provides more evidence of the limits of today's LLMS in interpreting humor shaped by distinct societal norms.

## 3.5 Summary and Research Gap

Prior studies on computational humor have made significant progress in monolingual humor detection, mainly in English, using either surface-level linguistic features or neural representations.

While humor detection has been extended beyond English to other languages, existing approaches largely treat humor as a language-agnostic classification task and do not explicitly account for culturally specific humor mechanisms. Newer efforts rooted in cultural context, such as the ManzaiSet dataset, demonstrate that understanding humor is deeply influenced by shared cultural knowledge and audience interpretation. Such differences become apparent when contrasting Western humor traditions, such as German joke formats, with Eastern approaches, such as Japanese wordplay and stage-driven comedy. This research typically examines people's reactions or tests models without explicitly probing culturally grounded mechanisms of humor. Still, tests on LLMs show mixed results in handling humor-related tasks, with indications of normative and cultural bias and open questions about their ability to generalize beyond English-dominated, Western-centric training data. Existing research has generally evaluated the humor of one specific culture/language (culture) and therefore is unclear if the humor model actually understands humor in other cultures or is simply identifying similarities/ patterns at the level of surface-linguistic features. Furthermore, most current benchmarks for evaluating humor are developed without explicit structural differences in humorous vs. non-humorous text. Therefore, it is possible that some humor models may achieve high performance by taking shortcuts based on structural characteristics of the text rather than a true understanding of the semantic content. To address these limitations, this dissertation will include a cross-cultural analysis of humor in German and Japanese and utilize a multi-tier framework. The first three tiers of the framework will establish baseline detection, rating, and explanation patterns. Subsequently, two new adverse testing tiers will be added to determine whether the model's ability to perform well would continue to exist even as controls are structured and semantically similar to those of jokes.

# 4 Research Question and Objectives

The main research question addressed in this thesis is:

> *Can large language models detect and interpret humor across German and Japanese, and what do their errors and explanations reveal about cultural understanding?*

From this, several sub-objectives follow:

- To develop multilingual humor datasets (Japanese and German), that allow for a cross-cultural comparison of humor.
- To compare humor detection performance of LLMs by employing structured metrics beyond binary classification (humorous / non-humorous; offensiveness; clarity; cultural awareness).
- To analyze qualitative differences in model behavior across language and humor types (wordplay, taboo, situational etc.).
- To determine whether strong humor-classification performance reflects genuine semantic and cultural understanding or reliance on surface-level textual cues.
- To test model robustness in increasingly controlled adversarial environments through the use of structurally equivalent non-humor examples and semantically 'dehumorized' counterfactuals—original jokes systematically stripped of their comedic elements.

## 4.1 Hypotheses and Expected Outcomes

Based on the theoretical background and related work discussed in Sections 2 and 3, this thesis formulates the following hypotheses regarding the cross-cultural humor understanding capabilities of LLMs.

**H1: Binary humor detection performance will remain high across languages.** LLMs are expected to achieve strong performance in Tier 1 binary humor classification for both German and Japanese datasets. Prior work has shown that transformer-based models can reliably distinguish humorous from non-humorous text by exploiting surface-level cues such as lexical incongruity, sentiment shifts, and atypical word associations, even in non-English settings. As the Tier 1 task does not require explicit cultural reasoning, performance differences across languages are expected to be limited.

**H2: Qualitative humor evaluation will show reduced consistency for culturally specific humor mechanisms.** In Tier 2, models are expected to exhibit greater variability in qualitative ratings, particularly on dimensions of clarity and funniness, when humor relies on language-specific mechanisms such as phonological wordplay or culturally embedded pragmatic conventions. In particular, Japanese dajare and structurally constrained comedic forms are expected to receive lower clarity and funniness scores than German jokes relying on semantic incongruity or irony, reflecting the models' reduced sensitivity to humor mechanisms that are not well represented in their training data.

**H3: Explanation-based evaluation will reveal limited cultural grounding.** In Tier 3, LLMs are expected to rely primarily on general humor descriptors, with limited and inconsistent reference to culture-specific linguistic or pragmatic factors. Explanations for Japanese humor are expected to be less precise and more generic than those for German humor, reflecting the underrepresentation of non-Western comedic traditions in model training data.

**H4: Apparent humor understanding will correlate with general reasoning ability rather than cultural competence.** Consistent with recent benchmark-based findings, models that perform well on explanation-based humor tasks are expected to do so primarily by leveraging general reasoning and language coherence capabilities, rather than by demonstrating genuine cultural understanding. As a result, high-quality explanations may coexist with misinterpretations of culturally grounded humor mechanisms, particularly in the Japanese dataset.

**H5: Performance will decline when non-humorous controls are structurally matched to jokes.** A decrease in overall performance of joke classification is expected when Humorous control texts are matched with neutral control texts that maintain dialogue style and structure. The decline in classification performance for these types of control texts will likely be due to an increase in the number of false positive classifications of the structurally matched control texts relative to a decrease in classification accuracy regarding the original joke.

**H6: Performance will decline further on semantically dehumorized minimal pairs.** We can anticipate that models will frequently classify the rewritten (non-humor) versions of original jokes (counterfactuals) with preserved setups, dialogue structures, and lengths, but without a punchline, as being humorous. This is because if models truly comprehend the humor mechanisms, they should be able to identify the difference between the originals and their dehumorized versions. Therefore, a model that produces a large number of false positive when evaluating these control versions would provide evidence that models do not recognize the removal of the semantic source of humor.

Overall, these hypotheses suggest that LLMs will perform strongly in detecting humor using traditional humor detection benchmarks. However, when tested in an environment of control and/or adversarial testing; their performance should decline. Therefore, we expect to see a divergence between success in the benchmarks used for humor detection and actual understanding of humor.

# 5 Methodology

This chapter describes the construction of the German and Japanese humor corpora, the evaluation pipeline used to assess LLMs across different levels of humor understanding, and the selection of models included in the study. The approach for humor evaluation is based on a three-tier framework introduced in Álvarez Jorge's paper [AJ24], focusing on humor classification, qualitative rating, and explanation. Following the initial results of Tiers 1-3 demonstrating unexpectedly high classification scores, the evaluation framework was expanded with two additional adversarial tiers to determine if the apparent success was due to a true understanding of humor or an ability to exploit structural features found within the data. The final methodology is therefore comprised of five tiers ranging from standard benchmark-style evaluation to controlled tests of semantic and cultural understanding.

## 5.1 Dataset Construction

The multilingual humor corpus consists of two components: a German humor dataset constructed through automated web scraping, and a Japanese humor dataset centered around a large-scale Japanese pun (*dajare*) corpus described in Yatsu and Araki [YA18]. All datasets were compiled using custom Python scripts that implement source-specific scraping, text extraction, cleaning, and filtering procedures. Such filters included: character length, word count, and dropping URLs, Emojis, excessive punctuation, and minimal humor markers ("lol", "笑", "haha") to reduce superficial cues that might bias classification.
To enable binary humor classification, each humor dataset is paired with a corresponding neutral (non-humorous) control corpus. Neutral texts were sampled from Tatoeba, a large, open, community-driven multilingual corpus consisting of natural-language sentences contributed by native and proficient speakers of hundreds of languages, including German and Japanese [Tat26]. Sentences from Tatoeba are intended to capture everyday language use rather than to convey expressive, humorous, or persuasive intent. As such, it provides a good basis for the neutral control data. It should be noted that all neutral sentences are authored by humans, therefore eliminating potential artifacts associated with synthetic or heuristically generated text.

The neutral sentences were collected separately for German and Japanese using language-specific filters and with a fixed sample size to provide equivalent class distributions for both languages. The neutral sentences were subject to a normalization and cleansing process, involving Unicode and whitespace normalization, character length limits, deduplication, and further filtering to eliminate fragments and formatting artifacts. A similar normalization and cleansing process was applied to the humorous texts, specifically targeted at removing noise caused by scraping and malformed items produced during the scraping process. By applying this type of processing, the possibility of models relying on surface-level features (e.g., format and extreme differences in length) decreases; thus, the differences in model performance can better be attributed to humorous content. Following the normalizing/cleansing processes, the humorous and neutral samples were then combined into uniform language-specific datasets using a common schema to combine both languages.
These Tatoeba-derived non-humorous control corpora were used as the initial comparison sets for the earlier phases of testing. However, after observing the results of the first three tiers, a methodological concern arose when it became apparent that such controls could be easily distinguished from jokes on either stylistic or structural grounds. This concern consequently motivated a second phase of the methodology, in which the study was augmented by additional adversarial control data. In Tier 4, the controls were structurally matched to the joke corpora via the use of subtitles obtained from OpenSubtitles. In Tier 5, the controls were derived from the original jokes themselves but were semantically dehumorized.

### 5.1.1 German Humor Dataset

The German humor dataset was created using an automated web scraping script that extracted humorous text from two public joke websites: `witze.tv` and `hahaha.de`. Both of these websites are joke repositories containing numerous collections of short jokes categorized by theme. These sources were selected based upon their widespread topical scope, uniform format, and public accessibility, supporting the collection of naturally occurring humor used online.
Data collection implemented through the Python script utilized standard HTTP requests and parsing through the use of BeautifulSoup. This allowed for pagination handling, various HTML page layouts, and non-uniform formatting across pages. Source-specific and general heuristic cleaning rules were used in conjunction with custom selectors to remove all non-textual components such as menu bars, banner cookies, header labels, and ads. Textual content representing the joke was the only item retained. All entries with either missing, empty, or malformed textual content were removed before processing.
The resulting German humor corpus spans a diverse set of joke categories, including *Flachwitze*, *Wortwitze*, *Kinderwitze*, *Blondinenwitze*, *Schwarzer Humor*, *Arztwitze*, *Männerwitze*, *Erotikwitze*, *Beamtenwitze*, *Berufswitze*, *Schulwitze*, and *Tierwitze*. Duplicate texts were removed through the application of exact text matching during preprocessing. After filtering and de-duplication, the final German humor dataset contained 996 unique humorous examples.
For the purpose of creating a balanced binary classification setting, neutral German control texts were drawn from the aforementioned **Tatoeba German Dataset**. The extracted texts were then filtered to remove short fragments, overly long passages, and other items that could be interpreted as noise, thereby providing comparable length and structural characteristics of the humorously oriented texts. The entire scraping, deduplication, and filtering processes were executed programmatically to allow for the reproducible construction of the German corpus and to provide a basis for comparison

with the Japanese corpus.

### 5.1.2 Japanese Humor Dataset

The Japanese humor dataset is constructed from a combination of a curated academic pun corpus and multiple publicly available Japanese humor websites. The primary source is a large-scale Japanese pun (*dajare*) corpus originally constructed by Araki and colleagues and described in Yatsu and Araki [YA18]. This corpus provides a substantial collection of short pun-based jokes and serves as a linguistically well-defined core for the Japanese humor data.
In order to increase stylistic diversity and include additional humor forms beyond puns, the dataset is augmented with web-scraped joke texts collected from various Japanese humor websites. The sites include collections of general jokes, *nazokake* (riddle-like wordplay), *tonchi* riddles, and short funny anecdotes. The web data is scraped using a Python script, analogous to the script used for the German dataset creation, and has been categorized by humor types during the collection. Preprocessing steps ensured that all entries belonging to a predefined set of humor categories were retained, and any empty or malformed entries were discarded. Only textual content was retained; user metadata and formatting elements were removed. A further challenge in constructing the Japanese dataset concerns the cross-linguistic comparability of short-form humor. Because some Japanese humorous forms are more interactional and context-dependent (Takekuro, 2006, as cited in [NR17]), certain types of humor may be less likely to appear in short, decontextualized textual formats that are directly comparable to German or English joke corpora. This may have contributed to the relative underrepresentation of provocative or taboo-oriented short-form material in the present dataset.
Although some of the sources included interaction signals such as upvotes and ratings, these were not used in this study. Instead, all of the humorous texts are considered equally and are treated in the same way, to be able to use them in conjunction with the German dataset and to prevent the combination of different annotation methods. Like the German dataset, neutral Japanese texts are selected randomly from the Tatoeba Japanese Corpus using the exact same cleaning and filtering process as the German data. This design ensures that cross-linguistic comparisons reflect differences in humorous content rather than dataset construction artifacts.

### 5.1.3 OPUS/OpenSubtitles Neutral Controls for Tier 4

An additional control dataset was created to serve as a non-humorous counterpart to the Tier 4 humorous dataset. The new non-humorous dataset was derived from the OPUS OpenSubtitles corpus, a large multilingual subtitle collection extracted from movie and television subtitles, provided from Opensubtitles, and distributed and translated through the OPUS project [LT16, Ope26]. The choice of using OpenSubtitles corpus as the source of the non-humor dataset was due to the stylistic similarity it provides to conversational exchanges (i.e., short conversational turns) as well as informal phrase structures, interruptions, and punctuation marks, all features which may mimic a joke's setup/punchline structure but are not intended as humorous. This therefore made the subtitle corpus appropriate to use as a source for generating non-humorous datasets that have similar structural and discursive characteristics as jokes.
The Tier 4 non-humorous dataset was generated using a Python pipeline (`Tier4_neutral_creation.py`) that is based on feature-matched retrieval instead of random sampling. Due to the fact that subtitles

are usually written as a series of short lines, the `Tier4_neutral_creation.py` script applied a sliding window concatenation technique. Subtitle lines were read sequentially and dynamically combined into text chunks so as to produce more organic multi-clause utterances resembling the setup structure of jokes. However, once a growing chunk reaches a predefined maximum size, the oldest line from the buffer is discarded. All metadata were removed using regular-expression-based cleaning rules. Any subtitle chunk containing explicit markers of humor (e.g. "haha", "LOL") were removed from the neutral control dataset. For the Japanese dataset, a whitelist of characters was implemented, which limited acceptable substrings to those valid for the Japanese writing system. Additionally, candidates were required to include at least some Hiragana to avoid accepting substrings composed entirely of Kanji or Katakana. Finally, density constraints were added to reject strings with high densities of whitespace or punctuation.

The script created a structural feature representation of all subtitles as well as all joke entries. The structural features used included character count, word count, sentence count, the number of exclamation marks and question marks and colon and ellipsis, the presence of quotation marks, and whether the entry begun with an interrogative cue. The process of matching was next conducted utilizing a nearest neighbor method. The distance score between the structural representation of each joke and its corresponding structural representations of each candidate subtitle segment was calculated. The distance function weighted differences in the number of words, sentences and punctuation usage. Each joke was then paired with the subtitle segment that resulted in the minimum structural distance. The resulting pairing produced a non-humorous control text that most closely approximated the visual appearance of the original joke.

Given that many jokes have very short lengths, the script imposed a constraint on the length of the non-humorous control texts. To reflect the prevalence of short punchline-like jokes, 45% of the non-humorous controls were constrained to fall within a short character-length range, while the remaining controls were selected through the broader matching procedure. The final output was a collection of 1,000 structurally similar non-humorous control texts. The control texts were saved in the same format as the humorous texts and were labeled as non-humorous controls.

The OPUS-based control set described above was the sole source of structurally similar controls used in Tier 4 as an alternative to the initial Tatoeba-based neutral control sets.

#### 5.1.4 Semantically Dehumorized Controls for Tier 5

The Tier 5 dataset was constructed through a four-step Python pipeline. The first step used a sampling script (`sample_tier4_jokes.py`) to create a controlled subset of humorous texts from the German and Japanese humor corpora. To assure the representative selection of various categories of humor, the script used a stratified sampling procedure by category with redistribution. Under the described procedure, each category of humor has been assigned a quota of samples; if there are not enough valid examples of a certain category, the missing part of the quota has been redistributed among all existing categories. A fixed random number generator has been used to make sure that the results can be reproduced.

Following the sampling of the humorous texts, we isolate the models' semantic comprehension by introduction a process we term 'dehumorization'. We define this as the systematic rewriting of sampled jokes into non-humorous counterparts. We implement this through the `gemini-2.5-pro` model and language-specific few-shot prompts in German and Japanese. The dehumorization is split into two methods: the "Mundane" transformation, and the "Minimal-Pair" transformation.

This was done in order to pinpoint exactly where an LLM's comprehension breaks down.
The "Mundane" transformation is implemented through `tier5_dehumorize_with_gemini.py`. The prompting strategy instructs the model to maintain the general topic of the original text whilst replacing the humorous punchline, wordplay, or absurdity with the factually correct and literally stated continuation. To ensure output quality, the script has been provided with a retry system having several guardrails. Firstly, after generating the non-humorous counterpart, the similarity between the original and generated text was measured. If the generated text was still too similar to the original text, a stricter prompt was issued. Secondly, the structural properties of the generated text have been compared with the structural properties of the original text. If the generated text had significantly changed in comparison with the original text in terms of either the number of sentences or total length, a new fallback prompt has been given to the model in order to enforce the stricter structural limitations. As a result of the second stage of the pipeline, a paired dataset has been obtained containing the original jokes and their non-humorous versions.
`tier5_dehumorize_min_pair.py` implements a more adversarial control condition, creating the non-humorous controls through a "Minimal-Pair" transformation. This aims to create the non-humorous counterparts as close as possible to the original joke in wording, setup, structure of dialogue and approximate length. The prompts used to instruct the model to keep the original setup and to modify only the element which generates the humorous effect. In addition, the prompts also ask the model to neutralize the explicitly stereotyped subjects when necessary (for example, to replace the demographic labels with more generic referents). This allows the resulting controls to check how far downstream classifiers will continue to interpret a text as a joke simply due to the fact that the text contains a known joke template or a stereotype cue. Due to the fact that the minimal-pairs are intended to be very similar to the original texts, at this stage, we have used even more strict thresholds to evaluate the similarity between the original and generated text before triggering the retries. We have also applied the same structural fallback procedures to maintain approximately the same number of sentences and length. Language-specific few-shot examples used for both the mundane and minimal-pair dehumorization prompts are provided in appendix section A.3.1 and A.3.2. The prompts used can be found in appendix section A.4.1 and A.4.2.
Thus, as a result of the Tier 5 dataset construction process, a set of the paired matches has been obtained, in which each original humorous text has been associated with one or more semantically neutralized non-humorous counterparts.

## 5.2 Evaluation Pipeline

The final evaluation design consists of five tiers. Tiers 1–3 are formed from the original evaluation structure and measured baseline humor classification, qualitative assessment, and explanation-based reasoning. Following an initial analysis of the first three tiers, two additional tiers were introduced as a form of adversarial extension to determine whether models would continue to perform well under conditions where structural and semantic shortcuts were systematically removed. All five tiers represent a progression from standard benchmark evaluation methods to increasingly restricted evaluations of true humor understanding.

#### 5.2.1 Tier 1: Binary Humor Classification

Tier 1 evaluates whether a model can distinguish humorous from non-humorous texts. Each model receives a single short text and must output a structured JSON response containing four fields: (1) a binary humor decision, (2) a confidence rating on a five-point scale, (3) a self-report on whether the model "finds" the text funny, and (4) a brief justification. The prompt follows the annotation format introduced in [AJ24], adapted to support both German and Japanese inputs.
Classification performance is computed against ground-truth labels stored in the merged corpus. Tier 1 does not involve any contextual information; each text is evaluated in isolation.

#### 5.2.2 Tier 2: Qualitative Humor Rating

Tier 2 extends Tier 1 by eliciting multi-dimensional qualitative ratings for each item. The aim is to examine how the models evaluate humor beyond the binary classification found in Tier 1. These ratings allow comparison across languages, humor categories, and models, revealing systematic differences in perceived funniness, clarity, originality, offensiveness, and appropriateness.
This tier does not require additional data collection and builds directly on the Tier 1 prompts.

#### 5.2.3 Tier 3: Explanation-Based Cultural Evaluation

In this tier, we conduct a qualitative analysis of the explanations produced by models during Tier 1 humor classification. Although models are not explicitly prompted to provide cultural analyses, their short justifications offer insight into the reasoning patterns they employ when judging humorous content.
Specifically, we analyze whether these explanations implicitly reference linguistic mechanisms (e.g., wordplay, incongruity), cultural assumptions, or social norms, and whether such reasoning differs across languages and error types.
Whereas Tier 2 evaluates cross-lingual consistency at the level of model outputs, Tier 3 focuses on the qualitative properties of the explanations accompanying these decisions.

#### 5.2.4 Tier 4: Structurally Matched Controls

Tier 4 tests if a model can continue to distinguish humorous and non-humorous texts when the control (neutral) texts have been structurally constructed to be similar to the humorous texts. The OPUS/ OpenSubtitles based control texts discussed in section 5.1.3 will be used. Non-humorous texts were selected such that they contain the same attributes as their humorous counterparts, decreasing the difference in number of dataset artifacts. The purpose of this tier is to test whether strong earlier performance reflects genuine humor recognition or reliance on surface-level structural cues.

#### 5.2.5 Tier 5: Semantically Dehumorized Controls

Tier 5 evaluation examines whether the model is able to distinguish humorous from non-humorous texts, when the control texts are developed by taking the direct content from the original jokes. In this tier, the original texts which are humorous will be paired with their semantically dehumorized counterparts (controls). The "mundane" and "minimal-pair" controls, previously discussed in

Section 5.1.4, will be tested separately, each paired with the corresponding original counterpart. The controls will eliminate all the semantic elements present in the original joke which were responsible for the original humor. This tier is consequently used to assess whether earlier success in humor classification reflects true comprehension or only the recognition of familiar humorous templates.

## 5.3 Models and Systems Used

The study evaluates a diverse selection of contemporary LLMs representing different families and training paradigms:

- **OpenAI model (GPT-4o-mini):** GPT-4o-mini was selected as the reference baseline model. As a compact and widely deployed member of OpenAI's GPT-4 family, it is designed for fast, low-latency inference on focused tasks and supports structured text outputs [Ope24]. This matches the present evaluation setting, which requires models to provide explicit natural-language explanations in addition to binary humor judgments. Prior work shows that instruction tuning with human feedback improves how well model outputs follow user intent and increases the quality of user-facing explanations [OWJ+22]. GPT-4o-mini therefore provides a practical and representative baseline for explanation-based humor evaluation while avoiding reliance on the largest proprietary variants.

- **DeepSeek model (DeepSeek-Chat):** DeepSeek-Chat was selected as a reasoning-oriented alternative to proprietary baseline models. Developed as a conversational system optimized for efficient inference, DeepSeek-Chat emphasizes general reasoning and multilingual capability. Prior work has shown that models trained under compute-efficient and reasoning-focused paradigms can exhibit strong transfer across downstream tasks without relying on maximal scale [HBM+22]. However, as a model developed under different regulatory and cultural constraints, DeepSeek-Chat may exhibit limitations in interpreting culturally specific or sensitive humor. Its inclusion therefore allows examination of how reasoning-oriented models trained under distinct cultural and institutional contexts respond to humor across languages.

- **X.AI model (Grok-4-1-fast-reasoning):** Grok-4-1-fast-reasoning was selected to explore the impact of alternative training data mixtures and stylistic biases on humor comprehension. Developed by xAI, Grok is trained on a diverse corpus that includes conversational and real-time web data, with an emphasis on contextual and informal language use. Prior research demonstrates that the composition of training data can substantially influence model behavior, even when architectures and prompts are held constant [GGS+20]. As a proprietary model with comparatively relaxed moderation constraints, Grok may be more permissive in engaging with humor that relies on social transgression or implicit contextual cues. Its inclusion enables analysis of how data provenance and moderation regimes shape humor interpretation.

- **Meta model (Meta-Llama-3.1-8B-Instruct):** Meta-Llama-3.1-8B-Instruct was included as an open-weight, capacity-constrained baseline. The Llama family has demonstrated that relatively small models can achieve competitive reasoning performance when trained efficiently on large-scale public data [TLI+23]. The choice of the 8B variant, rather than larger 70B-scale models, is deliberate: scaling-law analyses indicate that larger models are not always optimal under fixed inference or compute budgets, and that qualitative behavior can often be studied

effectively at smaller scales [HBM+22]. As an openly accessible model with known architecture and training composition, Llama provides a transparent point of comparison for assessing whether humor comprehension failures persist under reduced model capacity.

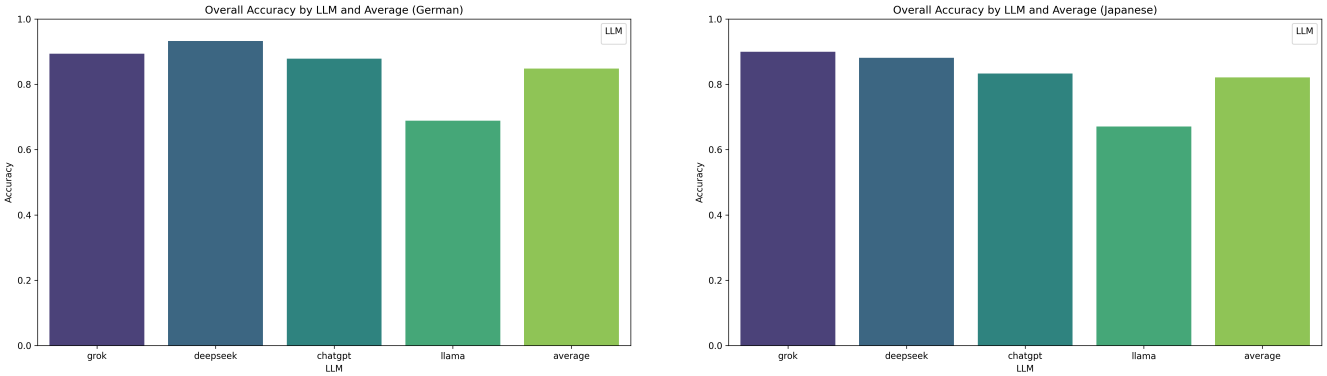
In addition to the previously described model evaluations, `gemini-2.5-pro` was used to support the creation of the Tier 5 dataset, producing semantically dehumorized versions of control texts. Gemini was chosen for this data preprocessing task because of the strict semantic and structural constraints required for the dehumorization. As such, the Tier 5 dataset required a high-capability reasoning model capable of tightly following instructions [CBS+25]. Given the role Gemini plays in the construction of the dataset, using a separate model for control generation reduced circularity in the benchmark design and minimized the risk that a model would later be evaluated on adversarial controls produced by itself or its own model family.
All evaluated models were prompted through a unified Python pipeline that ensured identical task instructions, output formats, and post-processing across APIs.

## 5.4 Ethical and Licensing Considerations

All data used in this thesis were obtained from publicly accessible web sources intended for entertainment or educational use. No private or user-generated personal data were collected. Web scraping respected the `robots.txt` guidelines of each domain and maintained minimal server load through rate-limiting. Attribution of sources is retained in the `source` column for each entry, ensuring transparency of provenance. For Tier 4, subtitle-based control texts were derived from the OPUS OpenSubtitles corpus in accordance with its citation and attribution requirements. For Tier 5, semantically dehumorized controls were generated solely for evaluation purposes from already collected humorous texts and did not involve personal or private user data.

## 5.5 Summary

To summarize, the methodology creates a reproducible pipeline for data collection, processing, cleaning, and structuring humor data in both German and Japanese for cross-cultural evaluation. The final design combined a three-tier framework for classifying humor, qualitative assessment, and explanation-based analyses, with two additional adversarial frameworks to assess whether models remain effective at evaluating humor despite structural or semantic shortcuts. As such, all five tiers will evaluate whether an apparent ability to recognize humor reflects actual semantic and cultural understanding.
The complete codebase, including all scraping scripts, preprocessing pipelines, evaluation code, and dataset construction procedures, is publicly available at `https://github.com/kaiepp/Bachelor-Thesis` [Epp25].




(a) German dataset

(b) Japanese dataset

Figure 1: Overall Tier 1 humor detection accuracy by model for German and Japanese datasets.

# 6 Results and Interpretations

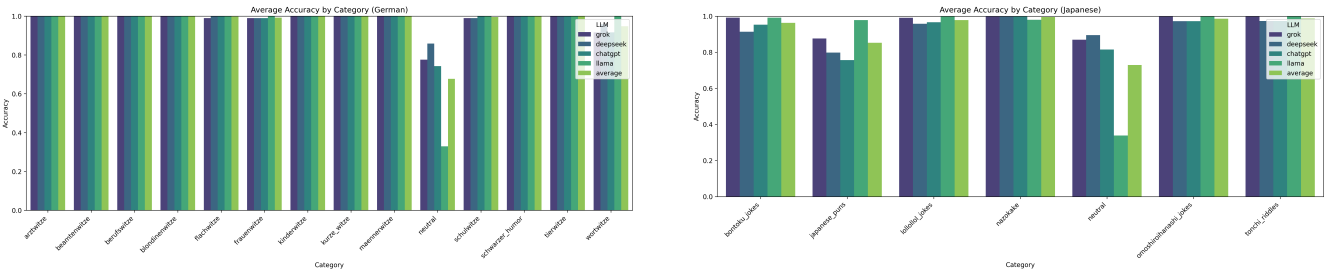
## 6.1 Tier 1: Binary Humor Classification

### 6.1.1 Results

Tier 1 examines whether LLMs are able to *identify* humor through a binary classification setting (humorous vs. neutral). This is evaluated for both German and Japanese languages, and under the four aforementioned models, using four complementary views: 1. Overall Accuracy, 2. Accuracy by humor subcategory (including the neutral control), 3. Model-reported self-confidence score, and 4. Mismatch percentage as an error-focused indicator.

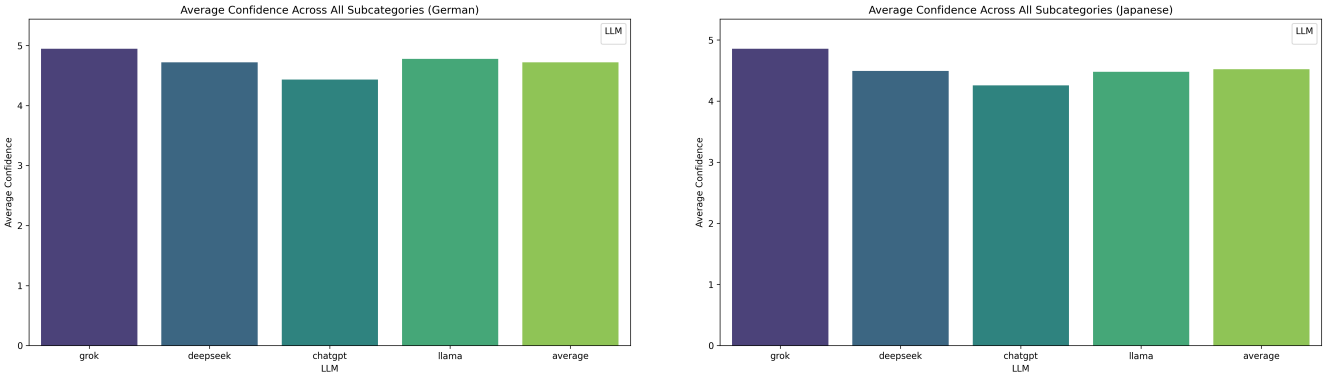
**Overall classification performance**
Figure 1 depicts the overall Tier 1 accuracy among the four models. This indicates the percentage with which the models were able to correctly classify humorous texts as humorous and vice versa. Across both languages, figures 10a and 10b, all models are seen to have achieved high accuracy, with stable differences between models that replicate across German and Japanese. This pattern suggests stable model-specific differences under identical prompting and evaluation conditions.

**Accuracy by humor subcategory and the role of neutrals**
In order to isolate the performance differences found in Figure 1, Figure 2 reports the accuracy of each model by humor subcategory, including the neutral control class. As illustrated in Figure 2a, German **accuracy is near the ceiling amongst all categories**, whereas the **neutral category exhibits the lowest accuracy**. As a result, the overall error rates are disproportionately driven by the false classification of neutral controls, rather than by the failures on canonical joke types. This demonstrates difficulties in correctly identifying the *absence* of humor in natural-language text and, by contrast, shows a strong ability in identifying humorous cues in jokes. Because the neutral class is intended as a strict non-humorous baseline, its comparatively lower accuracy is particularly informative for evaluating boundary cases between humor and non-humor. In Japanese, further difficulty in humor classification is illustrated through the decrease in accuracy of the *Japanese_puns* category. This category largely represents the Dajare, phonetic puns. This is consistent with the fact that its humor signal depends on sound-based similarity that is harder to recover from short text alone.




(a) German dataset (b) Japanese dataset

Figure 2: Tier 1 accuracy by humor subcategory (including neutrals) for German and Japanese datasets.




(a) German dataset (b) Japanese dataset

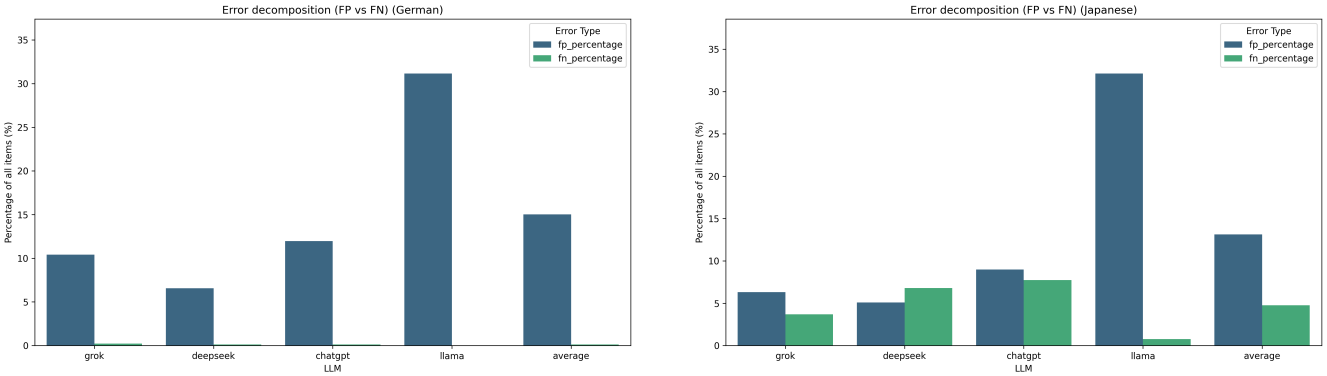
Figure 3: Average Tier 1 self-reported confidence by model for German and Japanese datasets.

**Self-reported confidence**

Figure 3 reports the average self-reported model confidence across all evaluated items on a Likert scale from 1 to 5. This represents the model's certainty in correctly classifying the text as humorous or not. Patterns reveal differences across the models' *calibration style*. The figure illustrates consistently high confidence scores across both languages and all models, with near symmetric scores between German and Japanese, with Grok scoring near ceiling scores at 4.95, and ChatGPT exhibiting the lowest score, whilst still scoring a high confidence score at 4.43. In spite of challenges previously seen with the comparatively difficult neutral category, along with high-confidence misclassifications (Found in Tier 3), high self confidence scores result in the notion that confidence should not be interpreted as a direct proxy for correctness; instead serving as an additional behavioral signal that can be compared against observed errors (Section 6.1.1).

**Classification error decomposition**

Accuracy alone is an incomplete metric in showing *how* models fail, as it collapses distinct error types into a single score. Figure 4, therefore, decomposes the overall error rate into false positives (FP) and false negatives (FN), reporting the classification error decomposition across all subcategories, making model-specific error rates and cross-language differences visible even with high accuracy scores. FP are created when neutral items are incorrectly predicted as humorous, while FN are created through incorrectly predicting jokes as non-humorous. While FP is noticeably dominant




(a) German dataset

(b) Japanese dataset

Figure 4: Tier 1 classification error decomposition by model for German and Japanese datasets.

across both languages, German shows near-zero FN, while FN is noticeably increased in Japanese; however, still FP dominant. Llama exhibits a significantly higher FP rate, at around 32.16% for the German dataset, and 32.13% for the Japanese dataset. This indicates a systematic bias towards predicting humor in ambiguous or neutral texts. Conversely, while the rest of the models concentrate on similar low FP rates, Deepseek shows the lowest across both German and Japanese, at 6.55% and 5.1% respectively. At the category level, provided in the Appendix (16–17), neutral items account for almost all errors found, while FN in Japanese are primarily found in *Japanese_puns* (dajare), and in *Wortwitze* in German.

### 6.1.2 Interpretation

Tier 1 clearly shows that all four tested models are able to reliably distinguish humorous from non-humorous test texts and perform similarly in both German and Japanese. Performance was consistently strong in both languages, model rankings were broadly stable, and overall behavior was similar across German and Japanese. Tier 1 therefore provides evidence of a narrow form of cross-lingual robustness. However, this similarity in behavior should not be taken as direct evidence of culturally grounded humor understanding. Rather, Tier 1 indicates that the humorous and non-humorous items in the benchmark appear to be separable under the present evaluation conditions.

One major finding of Tier 1 is that the most informative errors occurred not within the joke categories, but within the neutral control condition. Across both languages, the models achieved near-ceiling performance on humorous categories, but produced a significantly higher number of FP on neutral items. These findings suggest that the models are highly sensitive to cues associated with humorous form, but less reliable at identifying the absence of humor in borderline or stylistically ambiguous cases. Consequently, Tier 1 accuracy should not be viewed merely as evidence that the models understand humor; instead, it is primarily evidence that the benchmark allows for a relatively strong separation between jokes and neutral controls. Tier 1 accuracy is therefore influenced not only by model capability, but also by the construction of the dataset.

Furthermore, the comparison between German and Japanese illustrates the limits of aggregate-level interpretation. Although the models performed similarly in both languages at the aggregate level, the category-level analysis indicated that the two datasets were not equally easy for the models in

the same way. For example, the models were less successful in detecting Japanese *dajare* than in detecting many of the German joke categories. Many of the German jokes in the benchmark relied on explicit lexical or semantic incongruity, making them easier to detect in a binary classification setting, whereas Japanese *dajare* lacks these structural cues as humor is produced through homophones that rely on phonetics, a feature difficult for models to pick up on, especially if translation is used for humor detection. Tier 1 therefore supports the conclusion that cross-linguistic parity in headline accuracy does not imply equal sensitivity to all forms of humor. Parity can mask meaningful differences in sensitivity to various categories of humor.

Moreover, the confidence results reinforce this interpretation. The models reported very high levels of confidence in both languages, although they still made mistakes, particularly in the neutral condition. Confidence in this context therefore, appears to reflect the strength of the model's internal cueing rather than the actual reliability of the decision. If a text strongly matches a recognizable joke pattern or stylistic template, the model may make a highly confident judgment regardless of whether that judgment is correct. Confidence should therefore not be treated as a calibrated indicator of uncertainty in this study. Instead, it should be viewed as one of several behavioral signals showing how decisively the model applies its internal humor heuristics.

In summary, Tier 1 presents clear evidence that LLMs are capable of performing binary humor classification effectively under shared multilingual evaluation conditions. At the same time, Tier 1 identifies an important limitation that motivates the subsequent tiers. Strong performance on headline metrics does not, by itself, establish that a model genuinely understands humor or is culturally competent. Most errors occur in the neutral baseline, mechanism-specific joke categories such as *dajare* are harder to detect, and the models' confidence scores are saturated. Thus, the apparent robustness of the models is likely due in part to the separability of the benchmark and the models' reliance on surface-level textual cues. Tier 1 is therefore best understood as establishing a baseline level of classification performance.

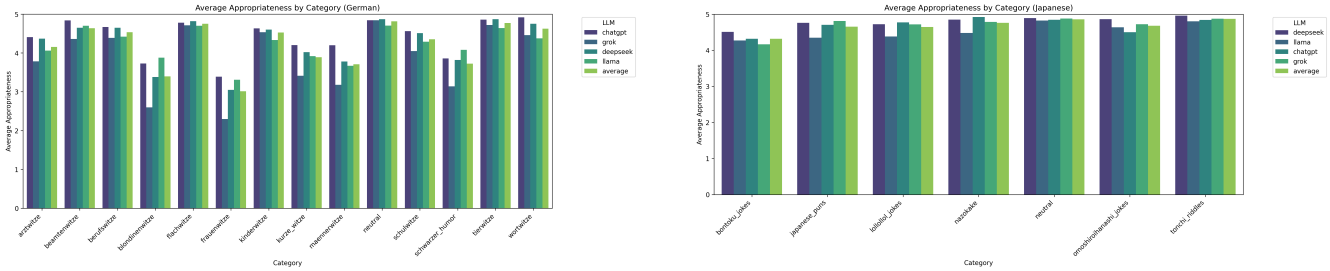
## 6.2 Tier 2: Qualitative Humor Evaluation

### 6.2.1 Results

Tier 2 presents a shift away from quantitative analysis, towards qualitative analysis, presenting an evaluation beyond binary classification of whether a text is humorous or not, and rather towards what kind of humor experience the LLM is attributed to each classification. This is implemented through an examination of graded judgements of *appropriateness*, *clarity*, *funniness*, *offensiveness*, and *originality*, on a five-point Likert scale. Importantly, this presents possible divergences in previous ratings, aimed at diagnosing what "robustness" really means.

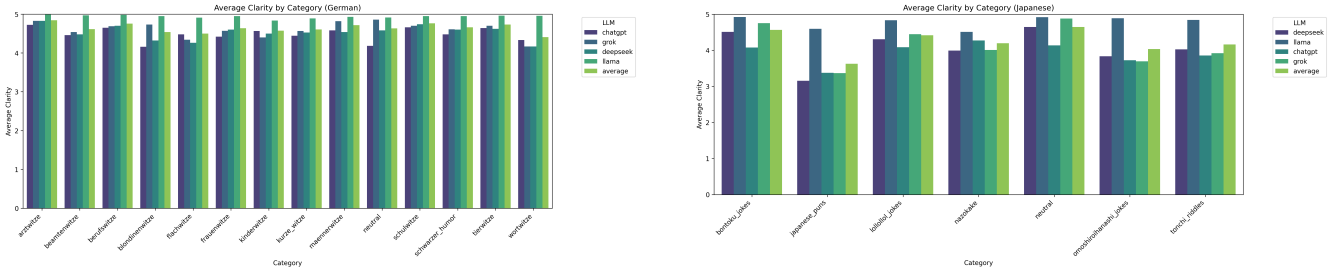
**Appropriateness by Category**

Figure 5 presents the average appropriateness scores for all four models, by humor category, for both German and Japanese. The general appropriateness scores are high across all categories. The German dataset exhibits stronger category-level heterogeneity, with categories such as *Blondinenwitze*, *Frauenwitze*, *Kurze_Witze*, *Männerwitze*, and *Schwarzer_Humor* consistently receiving lower scores than other categories, such as *Flachwitze*, *Wortwitze*, *Tierwitze*, and the *Neutral*, all categories that are 'low social risk' due to the lack of non-benign content such as profanity, aggression, or sensitive targets, prominent in the higher scoring categories. Japanese follows a similar trend; however, due to the use of anecdotal, semantic, and non-offensive jokes, the overall scores appear to be high




(a) German dataset (b) Japanese dataset

Figure 5: Tier 2 appropriateness ratings by humor category for German and Japanese datasets.




(a) German dataset (b) Japanese dataset

Figure 6: Tier 2 clarity ratings by humor category for German and Japanese datasets.

across all categories.

**Clarity by Category**

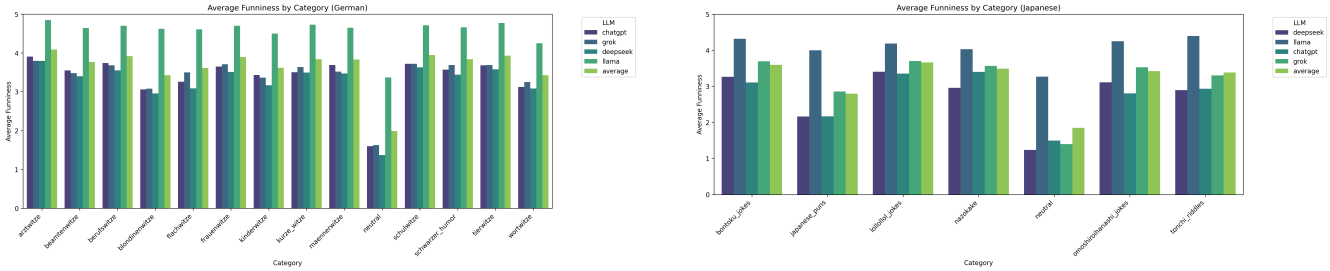
Figure 6 presents the Tier 2 clarity ratings by humor category for German and Japanese. Clarity scores are uniformly high across all categories in both languages, indicating that texts presented to the models are generally interpreted as understandable. German clarity scores exhibit the least variable with clarity scores consistently scoring above 4. Whilst also scoring high, Japanese shows stronger divergence among the category, especially for the *Japanese_pun* category, scoring notably lower with an average score of 3.62. This further substantiates the error rate seen in section 6.1.1 with the higher FP rates on the *Japanese_pun* category.

**Funniness by Category**

Figure 7 presents the funniness scores attributed to each text by all models for both languages. It reveals the greater dispersion in scores between models, with Llama consistently scoring the highest funniness scores for all categories in both languages, compared to the other models. This, along with the lower accuracy scores from Llama, reveals a lower humor threshold, leading to greater FP rates on neutrals. Neutrals are, however, perceived as the least funny in both languages, along with the *Japanese_pun* category, further supporting the previous notion of its difficulty for models to classify.

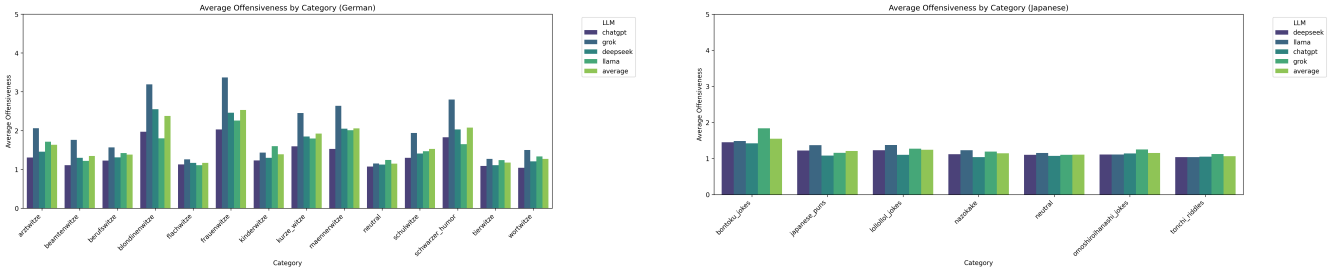
**Offensiveness by Category**

Offensiveness ratings found in Figure 8, depict clear patterns that have been previously established. These ratings exhibit strong inverse proportional ratings to the appropriateness ratings found in Figure 5, with German humor categories associated with stereotypes and social groups, such as




(a) German dataset (b) Japanese dataset

Figure 7: Tier 2 funniness ratings by humor category for German and Japanese datasets.




(a) German dataset (b) Japanese dataset

Figure 8: Tier 2 offensiveness ratings by humor category for German and Japanese datasets.

*Blondinenwitze*, *Frauenwitze*, *Maennerwitze*, and *Schwarzer_Humor*, exhibiting elevated offensiveness scores, while other categories remain low. Japanese scores remain uniformly low across all models and categories.

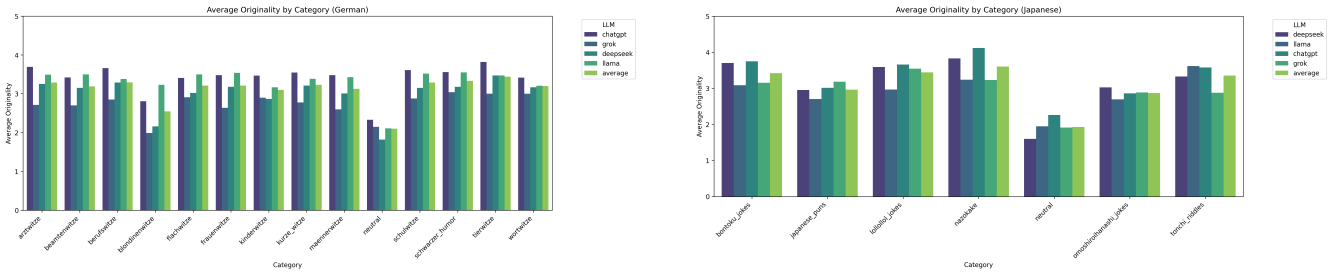
**Originality by Category**

Across models, the originality scores seen in Figure 9 are perceived to be moderate in both languages, clustering around the mean, with an overall originality mean of 2.64 for German and 2.51 for Japanese, indicating a conventional judgment towards most items. At the category level, neutrals exhibit the lowest scores for both languages, indicating that the models are able to somewhat distinguish non-humorous material as they are designed to be mundane. Notably, the relative ordering of the categories by originality differs, suggesting that originality judgements are not only influenced by the underlying content itself but may also reflect genre conventions and cultural familiarity.

### 6.2.2 Interpretation

Tier 2 builds on Tier 1 by adding a more fine-grained view of model behavior through its focus on how humor is evaluated, rather than simply classifying a given text as either humorous or not. While both sets of findings show that, at the aggregate level, the models provide very similar evaluations of humor in German and Japanese, there appears to be significant variation in how humor is judged across the categories within each language.

Category-level analysis provides a more informative picture than overall averages. For instance, while the average scores for the five aspects of funniness, clarity, offensiveness, originality, and




(a) German dataset

(b) Japanese dataset

Figure 9: Tier 2 originality ratings by humor category for German and Japanese datasets.

appropriateness, are generally quite comparable for the two languages, the by-category ratings reveal substantial variations in judgment between the two languages. This is particularly evident regarding aspects related to humor quality, where the Japanese categories have greater variability than the German categories. These results support a more limited definition of robustness. The models provide similar outputs under the shared evaluation framework, but this is not sufficient to claim that they evaluate all humor mechanisms equally well across languages.

Appropriateness and offensiveness ratings are especially informative in this regard. In the German dataset, categories containing stereotypes, social groups, or content that could be considered socially risky were consistently rated lower on Appropriateness and higher on Offensiveness than categories containing humor that was structurally simpler and socially less risky (e.g. *Flachwitze*, *Wortwitze*, *Tierwitze*). Categories rated more highly on Appropriateness, and lower on Offensiveness, included those containing jokes based on simple wordplay. The Japanese dataset provided a much flatter profile for these categories, as many categories received high ratings for Appropriateness and low ratings for Offensiveness. This difference is however attributable to dataset composition, as the Japanese corpus had fewer humor categories centered on provocative and socially sensitive content compared to the German dataset. As was discussed in Section 5.1.2 and Section 2.2, this may be a reflection of not only the source composition differences, but also the cross-linguistic comparison constraint. Some Japanese humorous content types are typically more interaction and context based, and are consequently less likely to exist in the form of short, decontextualized texts for direct alignment with other joke corpora such as the German dataset. The Tier 2 pattern therefore suggests that the models are attuned, at least to some degree, to differences in the social and normative profile of humor categories, but that the degree to which this occurs is influenced by the composition of the dataset. In short, Tier 2 has shown that evaluating the quality of humor is not just about determining if a text is funny, but also includes an assessment of whether the model views the text as socially acceptable, risky, or conventional.

Clarity and funniness ratings further refine interpretation of Tier 1 results. The ratings for Clarity are generally high, suggesting that the vast majority of texts are being interpreted as clear by the models. However, the Clarity ratings for the Japanese *dajare* category are notably lower than all other categories. This supports the findings from Tier 1 that this category is comparatively difficult to classify. While part of this pattern may reflect the classification noise found in the data, the lower Clarity rating for *dajare* suggests that the difficulties encountered in Tier 1 are not solely due to classification noise, but that there may be fundamental limitations to how the models process humor mechanisms that rely upon phonetic similarity, or linguistic structure. A similar trend is

evident in the Funniness ratings, where the lowest ratings for Funniness are assigned to Japanese *dajare*-based humor, and neutral items. This suggests that not only are the models less likely to classify such items as humorous, but also less likely to assign a high funniness scores to such items when evaluating them directly.

Moreover, differences in model behavior become more apparent in Tier 2. One notable example is Llama, which consistently evaluates categories as being more humorous than the other models. When combined with the fact that Llama had lower accuracy than the other models in Tier 1, and a higher FP rate on neutral items, it is reasonable to conclude that Llama uses a lower threshold for determining that an item is humorous. Therefore, Tier 2 provides additional methods for distinguishing between models that may perform similarly in terms of overall accuracy, but use different approaches to determine whether an item is humorous.

In general, Tier 2 supports the broader conclusion that having strong multi-language performance on headline metrics does not guarantee that models will evaluate humor in the same culturally grounded manner. The qualitative ratings provide evidence that models can generate relatively consistent judgments across languages, but that this consistency is far from universal, and depends heavily on the type of humor, the type of social content, and the linguistic mechanism. As a result, Tier 2 demonstrates that cross-linguistic robustness is largely a product of aggregation. If one examines the models' judgments, based on the type of humor, social content, and language-specific mechanism, then one finds that the judgments produced by the models differ. Therefore, the apparent robustness of the models is best seen as a product of the models' behavior under shared task constraints, rather than as a demonstration of a deep understanding of cross-cultural humor.

## 6.3 Tier 3: Explanation-Based Error Analysis

### 6.3.1 Results

Tier 3 introduces a qualitative analysis that complements the quantitative metrics presented in Tier 1 by inspecting representative FP and FN through the use of model justifications of its humor classifications. The goal is to identify the causes of why the models fail, and whether they are caused by model limitations or dataset construction introduced artifacts. Tier 3 provides strong insights into the observed findings from the previous tiers.

**German (Tier 3).** Across the German samples, most FN are found to be rare, which is consistent with the mismatch percentage found in section 6.1.1. In total, across all four models, 11 out of 2151 items were FN (ChatGPT (3), Grok (5), DeepSeek (3), and none for Llama). While these cases are isolated, they provide valuable insight. In particular, a large number of FN in the German dataset were produced as a result of incomplete or degraded joke texts caused by issues with the scraping of the data. Manual examination of sampled cases shows that several of these FN were produced by joke texts being either truncated or having their punchline removed during the scraping process. As a result, the fragments of jokes remaining contained insufficient information to indicate the presence of humor, but were nonetheless correctly classified as non-humorous by the models. However, because the joke texts had been labeled as humorous by the original source metadata, the errors reflected an alignment problem between the ground truth labels provided and the actual textual input provided to the models, rather than an indication of the models failing to reason correctly. This case is illustrated by a Tier 3 exemplar error in Appendix C.3, where the initial fragment contained an HTML link and was removed during preprocessing, leaving only a

residual (non-humorous) segment while the source label remained "humor". In addition, because the source website operates as an open community forum rather than a curated joke archive, some labeled "jokes" are effectively unmoderated comments or low-quality submissions that are not clearly humorous in isolation.
By contrast, German FP follows a more consistent linguistic pattern. A large portion consisted of multi-sentence neutral texts that the model perceived as structurally humorous, which led to the following reasoning: "The text is a classic anti-joke setup, where the expected joke is subverted by a mundane, factual answer", as shown in Appendix C.1. These texts often exhibited discourse structures typical of jokes, indicating that slight cues are sufficient for models to classify the element as humorous. Additional FP arose due to humorous ambiguity in neutral texts. Neutral texts containing poetic or rhetorical tones produce signs of figurative language or mild irony, leading the model to falsely classify the element as humorous.

**Japanese (Tier 3).** In the Japanese sample, fewer instances of scraping degradation were observed, and errors more often reflected specific humor mechanisms. In particular, a large share of FN occurred for *dajare* (phonetic wordplay) jokes. By contrast, many true positives were anecdotal or situation-based jokes whose humor is more semantically explicit and thus more robust to paraphrase or translation. FP in Japanese again clustered in the neutral class, particularly for sentences containing ambiguity or expressions that allow multiple pragmatic readings depending on context.

**Model-specific observations.** A small subset of Japanese FP, especially in Grok, appeared to be triggered by surface features that resemble humor under a Western comedic framing (e.g., marked lexical choices or stylized orthography), even when such usage is conventional in Japanese. While this does not establish a systematic cultural bias, it is consistent with the hypothesis that some heuristics transferred from Western humor priors may over-trigger on stylistic or lexical features in Japanese.

### 6.3.2 Interpretation

Tier 3 provides some of the strongest support for the central thesis claim that strong classification performance does not always imply culturally grounded humor understanding. The Tier 3 sample of the explanations produced by the models demonstrates how models typically rely on general cues (e.g., wordplay, irony, unexpected contrast, exaggeration) to identify humor rather than understanding why the joke is funny. Therefore, the explanations generated by the models are generally plausible; however, they tend to be generic and applicable to humor in other languages and humor types. This pattern demonstrates that they have produced convincing humor-related output while relying upon shallow textual regularities rather than developing a deep pragmatic understanding of humor [JK23, N+25]. Thus, within the context of this thesis, Tier 3 supports the argument made in Section 2.3 that successful performance of humor benchmarks may represent the ability of the model to recognize structural cues related to humor, rather than a deep understanding of culturally embedded meaning.
One of the most significant findings is that the models' similar levels of aggregate performance in German and Japanese cannot be taken as evidence that the models interpreted humor in the same way in both languages. Rather, the fact that the models used the same explanatory templates in both languages indicates that the cross-lingual robustness that was observed is primarily the result of recurring textual patterns being represented at an abstract level. This is significant

because, as described in Section 2.2, humor is not simply a formal textual property, but a culturally situated linguistic practice that is influenced by background knowledge, pragmatic conventions, and language-specific mechanisms. Therefore, if models generate explanations for humor that do not reference these mechanisms, then the stability of the models' performance across languages is best characterized as the repeated application of a heuristic, rather than as a culturally grounded humor understanding.

The difference in the FN patterns between German and Japanese is especially interesting, as it shows that superficially similar errors can result from quite different sources. For example, in the German sample, many FN could be better explained as resulting from artifacts of the dataset rather than from a failure of reasoning. Specifically, the German sample contained truncated joke texts, removed punchlines, and residual fragments remaining after processing, all of which resulted in the loss of the humorous signal in the textual input, despite the original humorous label. Tier 3 thus suggests that at least part of the German FN pattern can be attributed to the construction of the benchmark itself, as opposed to failures by the models to identify the humorous signal. This finding supports the criticism expressed in Section 2.3 as well as prior work [MKTO25], that humor benchmark results are potentially distorted by structural characteristics of the dataset itself.

In contrast, the Japanese FN frequently reflect a genuine interpretative limit. Specifically, a large proportion of the Japanese FN involved dajare – a type of humor that relies on phonetic similarity as opposed to explicit semantic incongruity. As discussed in Section 2.2, Japanese humor often relies on phonological properties, social and interactional conventions, and subtle contextual factors. These features are difficult to derive from short written segments, and are especially difficult to recover when the humor relies on homophony. Tier 3 thus supports the conclusion drawn in Tiers 1 and 2 that the reduced performance on dajare represents a specific model limitation in identifying humor that is embedded in linguistic form, rather than in overt semantic content. In this sense, the Japanese data present stronger evidence of a genuine modeling limitation than do the German data.

Finally, the FP patterns provide additional insight into this interpretation. Two subsets of FP appear to arise from distinct sources. First, some FP appear to result from ambiguity inherent in the neutral control texts themselves. Many neutral texts include irony, rhetorical contrast, or mild ambiguity, making a humorous interpretation defensible. Second, some FP appear to arise from a more direct model-based source. In this case, the models apply humor heuristics learned during training too broadly, and attribute a humorous intent to texts that are not pragmatically intended to be humorous, including tragic or stylistically marked statements. The distinction between these two subsets is important, since it illustrates that FP are not a uniform phenomenon. Rather, FP can represent either instability in the boundary between humorous and non-humorous texts, or a model's tendency to generalize from surface-level cues too aggressively. Furthermore, the FP observed in Grok for Japanese may also indicate that certain stylistic or lexical properties are being interpreted via culturally misaligned priors. This is consistent with broader concerns that models trained on the dominant distribution of training data will tend to over-apply those patterns across cultures [SCNP21, PPJ+25].

Tier 3 also helps clarify the confidence findings of Tier 1. High-confidence errors arise when the model assigns a confidence rating of 4 or above. This shows an indication of how strongly the model matches the text to its learned humor mechanism. It appears that confidence in these instances is an indication of the strength of the cues the model relies on, and not the reliability of the classification. When the cues learned in training match the intended humor mechanism, the performance of the model is high, however, when the cues do not match the intended humor mechanism, the model can

still produce errors with high certainty. Consequently, confidence should be viewed as an indicator of how certain the model is that the cues it is using to make the prediction are correct.
Therefore, Tier 3 demonstrates that the errors made at previous tiers have two primary causes: artifacts related to the benchmark construction methods, and limitations of models capturing humor mechanisms from short texts that are specific to cultures and languages. In addition, Tier 3 illustrates the critical point that simply producing fluently articulated explanations does not equate to achieving a deep understanding. While the models frequently generate coherent explanations of their decisions, those explanations are generally generic, weakly culturally grounded, and influenced primarily by surface-level humor heuristics. As such, Tier 3 is the most compelling evidence supporting the central argument of this thesis: that although LLMs can achieve strong results on multilingual humor benchmarks, they are still far from achieving a nuanced understanding of humor as a culturally situated linguistic practice.

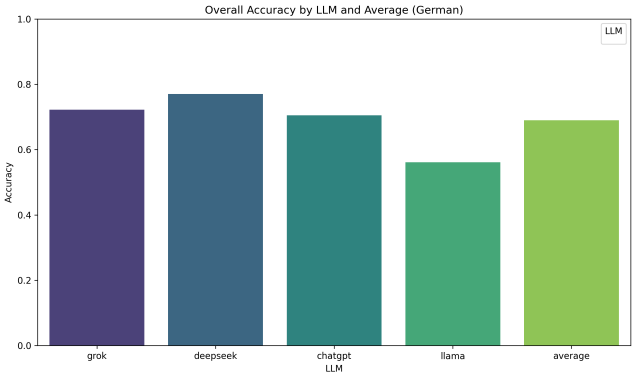
## 6.4 Tier 4: Structurally Matched Controls

### 6.4.1 Results

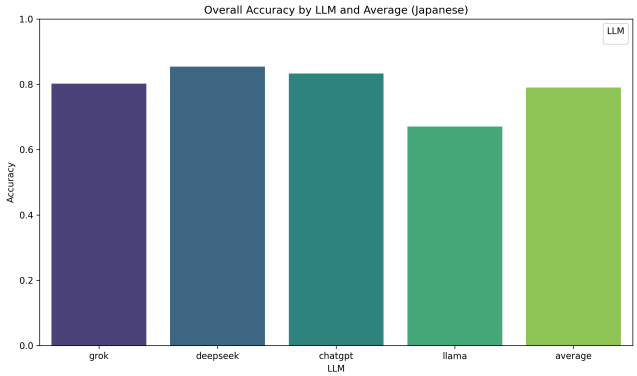
The Tier 4 evaluation assesses the extent to which models perform on humorous and non-humorous examples that have been structurally matched using subtitle data. Unlike previous non-humorous control items, which simply contained neutral examples of text, these items were designed to appear to be jokes at first glance while being non-humorous. These results are presented in two ways: first, each model's accuracy across the entire Tier 4 set (Figure 10), and secondly, each model's accuracy when divided into performance on humorous examples versus non-humorous examples that are structurally matched (Figure 11). The choice of visualization for figure 11 stems from findings previously found in section 6.1.1, where lower accuracy scores were primarily driven by FP, i.e., incorrect classification of non-humorous control items.

**Overall accuracy.** Tier 4's overall accuracy for each model is represented in Figure 10. As illustrated in the figure, we see that for all models in both languages, Tier 4 scores lower in overall accuracy than the humor dataset used in tiers. Similar accuracy patterns are found in both the German and Japanese dataset: DeepSeek has the highest overall accuracy, and is followed by Grok, and then ChatGPT. Llama consistently ranked much lower in accuracy than the other three models. We also observe that the average Tier 4 accuracy in Japanese was greater than the average accuracy in German, as shown in Figures 10a and 10b.

**Accuracy on jokes vs. matched controls.** We further visualize Tier 4 accuracy of each model through separation of input domain. As can be seen in Figure 11, the accuracy of each model varies significantly depending on whether the item presented to the model is humorous or not. Each model demonstrated extremely high accuracy when classifying the joke items. However, the accuracy of each model drops dramatically when attempting to classify the structurally matched non-humorous controls.
In the German dataset, all models classified the joke items with near-perfect accuracy; however, when evaluated on the structurally matched controls, the accuracy of the models decreased dramatically. When comparing the accuracy of the models on the structurally matched controls, DeepSeek achieved the highest accuracy and was followed by Grok and ChatGPT. Llama demonstrated the lowest accuracy when compared to the other three models (Figure 11a).

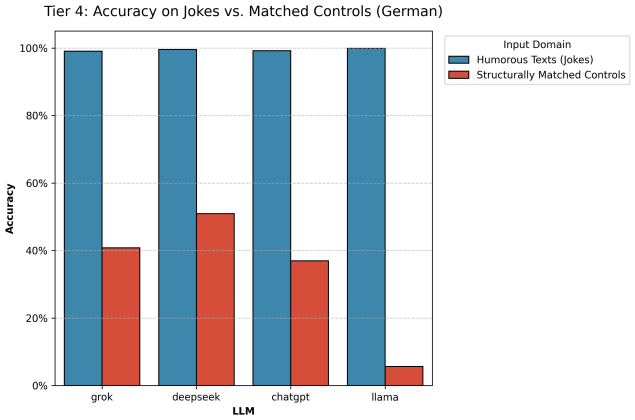



(a) German dataset

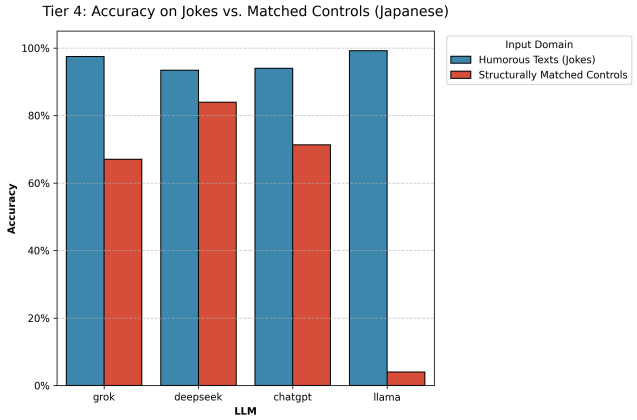



(b) Japanese dataset

Figure 10: Tier 4 overall accuracy ratings by model for German and Japanese datasets.




(a) German dataset




(b) Japanese dataset

Figure 11: Tier 4 accuracy by input domain for German and Japanese datasets, comparing humorous texts and structurally matched controls.

A similar pattern appears in the Japanese dataset. Accuracy on joke items again remains very high across all models, whereas control accuracy is consistently lower. DeepSeek achieves the highest score on the matched controls, followed by ChatGPT and Grok, while Llama again performs substantially worse than the other models (Figure 11b).

Overall, the Tier 4 results show a clear separation between performance on humorous texts and performance on structurally matched non-humorous controls in both languages. While joke classification remains near ceiling, the control condition accounts for most of the reduction in overall Tier 4 accuracy visible in Figure 10.

### 6.4.2 Interpretation

Tier 4 is the most straightforward test yet of whether performance on humor-classification tests reflects true humor recognition or reliance on superficial patterns in the text. When neutral controls are replaced with structurally identical controls, the previous near-ceiling results do not hold. The main insight is not that average accuracy drops, but that this drop occurs almost exclusively due

to a decreased ability to correctly classify non-humorous controls, while the recognition of the initial jokes continues to remain near-ceiling. This asymmetry is highly informative. If models were using semantic mechanisms to determine what made a text funny, the replacement of the earlier non-humorous controls with controls that resemble jokes should have only slightly impacted their performance. The large decline in accuracy when the controls are structured similarly to the jokes indicates that the joke-like surface structure of the text remains a significant signal for the model. This is the same methodological concern that was identified in Section 2.3: when humorous and non-humorous texts in a benchmark systematically differ in terms of structure, models will likely learn to exploit these structural differences rather than learning to identify humor itself. Therefore, Tier 4 directly supports the claim that a large portion of the high performance seen in the earlier tiers was supported by structural shortcuts, as opposed to actual humor understanding [JK23, MKTO25].

This interpretation is further supported by the dataset construction. As described in section 5.1.3, the OpenSubtitles-based controls were specifically designed to match the length, sentence structure, punctuation, and dialogue-like presentation of the jokes. Therefore, Tier 4 was not created to make the task arbitrarily harder; instead, we eliminated the stylistically distinct separation between jokes and the earlier neutral baseline. Therefore, the decline in control accuracy should be viewed as evidence that the models benefited from the relatively clear separation of the earlier benchmarks, rather than as evidence that the new dataset is inherently more difficult. In other words, Tier 4 reduces the gap between the forms of humorous and non-humorous inputs, and once the gap is reduced, a substantial portion of the apparent humor competence diminishes. This aligns well with prior work demonstrating that LLMs can provide plausible humor judgments based on textual regularities rather than pragmatic or culturally grounded understanding [JK23, N$^+$25].

In addition to providing further support for the claims of Tiers 1–3, it provides evidence that the models' ability to recognize humor cues was much greater than their ability to recognize the lack of humor in borderline cases. Tier 3 demonstrated that a large number of FP occurred in the neutral category due to the texts containing discourse patterns associated with anti-humor, irony, or conversational punchlines. Tier 4 has built upon this finding in a controlled environment: since the neutral texts were structured to resemble the jokes, the FP behavior became far more visible. Therefore, Tier 4 does not contradict the previous three tiers; rather, it provides an explanation for why the results of the first three tiers occurred. The high performance of the models in distinguishing between neutral controls and humorous texts in the first three tiers appeared to be due, in part, to the ease of distinguishing between the two based on stylistic grounds. Tier 4 serves as an important validation step for the central claim of this thesis: the high benchmark accuracy of a model is not sufficient to conclude that the model understands humor as a culturally embedded phenomenon.

The model-specific comparisons are also revealing. Specifically, while all models experienced a significant loss in accuracy when the controls became joke-like in structure, the degree of that loss varied across models. For example, DeepSeek was the highest-performing model when the controls were structurally matched, indicating that DeepSeek may be less reliant on joke format than the other models. Although DeepSeek still experienced a significant loss in accuracy, it was less severe than the losses experienced by Grok and ChatGPT. Both Grok and ChatGPT maintained high levels of accuracy when identifying actual jokes, yet lost a significant amount of accuracy when identifying structurally matched controls. This indicates that the earlier success of these models was partially due to their ability to identify familiar forms of humor. On the other hand, Llama exhibited the greatest loss in accuracy among the four models. While Llama's joke identification

accuracy remained near perfect in both languages, its control identification accuracy collapsed almost completely. This is particularly important because it illustrates that high levels of success in identifying humorous items do not preclude the possibility of failing to reject structurally similar non-humorous items. In fact, the nearly complete collapse of control accuracy in Tier 4 indicates that Llama appears to apply a permissive humor heuristic whereby it accepts a wide variety of joke-like structures as humorous. This is consistent with the Tier 2 results in which Llama scored higher on funniness scales, and with the Tier 1 results in which it generated the largest number of FP. However, it should be noted that this comparison is not entirely fair to Llama. As described in Section 5.3, the Llama variant used in this study is the 8B-parameter model, which is many times smaller in capacity than the other three models evaluated. At this scale, it is questionable whether the model performs consistent instruction following reliably across all items in the task, and some of the observed behavior, including the near-complete collapse of control accuracy — may therefore reflect general limitations in instruction adherence at reduced model capacity rather than a principled humor heuristic. This does not invalidate the comparison, as the 8B scale was a deliberate choice motivated by transparency and accessibility, as discussed in Section 5.3. Nevertheless, model-size differences should be kept in mind when interpreting the gap between Llama and the other three models, and the patterns observed in Llama are best treated as indicative rather than directly comparable to the proprietary, larger-scale models.

A further point of interpretation concerns the cross-linguistic comparison. The extent to which the shortcut effect is present in both German and Japanese implies that the reliance on structural cues is not confined to one linguistic or cultural environment. This finding is significant as the cross-lingual stability of model behavior does not indicate cross-cultural understanding, but may instead reflect the transfer of the same superficial heuristic across languages. On the other hand, while all four models show a decrease in accuracy on the Tier 4 controls compared to the Tier 1 controls, there was a somewhat greater drop for the German controls. This result must be viewed with caution. The fact that there were fewer errors for the Tier 4 controls in the case of the Japanese data does not imply that the models have a deeper understanding of Japanese humor. Rather, the findings suggest that the relationship between the humorous and structurally similar non-humorous texts is slightly different for each of the two datasets. In addition, the separation between the Japanese Tier 4 controls and jokes appears to be greater than that of the German controls for some models. Therefore, the overall trend of the results in terms of cross-lingual differences in Tier 4 is consistent with the broader argument in the related work, in which, multilingual humor evaluation may blur cultural understanding with the properties of the language-specific datasets. Thus, cross-lingual differences in Tier 4 are informative, but they cannot be used as a straightforward basis to rank how well the models "understand" the cultures' humor [BBBB19, WS19].

In conclusion, Tier 4 demonstrates that much of the robustness observed in earlier tiers is contingent upon the design of the benchmarks. If, in the benchmarking process, non-humorous texts are allowed to remain structurally distinct from joke texts, then the models' ability to detect humor appears to be quite strong. Once non-humorous texts are made structurally similar to joke texts, however, the models perform substantially worse. Furthermore, this poor performance is reflected in high rates of FP on the matched controls. Therefore, the primary implication of Tier 4 is not that the models completely fail at detecting humor, but rather that their success is more sensitive and heavily reliant upon surface-level patterns than suggested by headline accuracy scores. Finally, Tier 4 provides direct support for Hypothesis 5 and demonstrates the need for the even stronger semantic controls introduced in Tier 5.

## 6.5 Tier 5: Semantically Dehumorized Controls

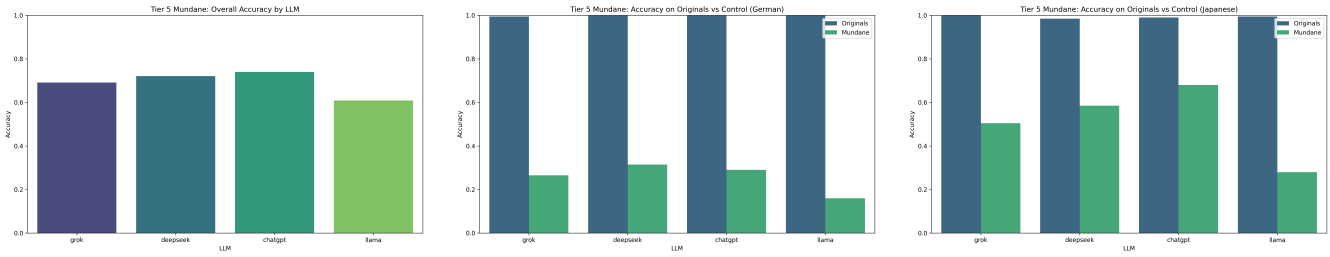
### 6.5.1 Results

Tier 5 assesses whether the models can continue to distinguish humorous from non-humorous texts when the control items are directly derived from the original jokes using semantic dehumorization. The results are presented separately for the *mundane* and *minimal-pair* transformations. Performance for each transformation is represented through overall accuracy by model across both languages, as well as accuracy broken down for the original jokes and the dehumorized control items in the German and Japanese datasets.

**Mundane transformation.** Figure 12 presents the Tier 5 results for the mundane transformation. Across both languages in the aggregated results, overall accuracy for all models was lower than in the previous tiers. The consistent relative ranking between the models remains, with ChatGPT, DeepSeek and Grok consistently forming a similar performance cluster, whilst Llama lags in performance.
When performance is separated by input type, all four models maintained nearly perfect accuracy on the original jokes. The accuracy of the mundane dehumorized controls is significantly reduced for all models. All models classified the original jokes in both languages with nearly perfect accuracy, while a sharp decrease in accuracy occurred for all models when classifying the mundane dehumorized controls. In the German dataset, DeepSeek achieved the highest accuracy in classifying the mundane controls. ChatGPT and Grok followed. The lowest accuracy among the German mundane controls was recorded by Llama. The same asymmetry persisted in the Japanese dataset. High accuracy remained for the original jokes; however, the accuracy of the dehumorized controls remained substantially lower. In the Japanese dataset, ChatGPT achieved the highest accuracy in classifying the dehumorized controls. DeepSeek and Grok followed. Once more, Llama recorded the lowest accuracy among the Japanese dehumorized controls.
Overall, the results of the mundane condition demonstrate a consistent division between the high accuracy of the models in identifying original jokes and the substantially reduced accuracy of the models in identifying dehumorized controls, with the accuracy of the German controls yielding lower accuracy than that of the Japanese controls for all four models.

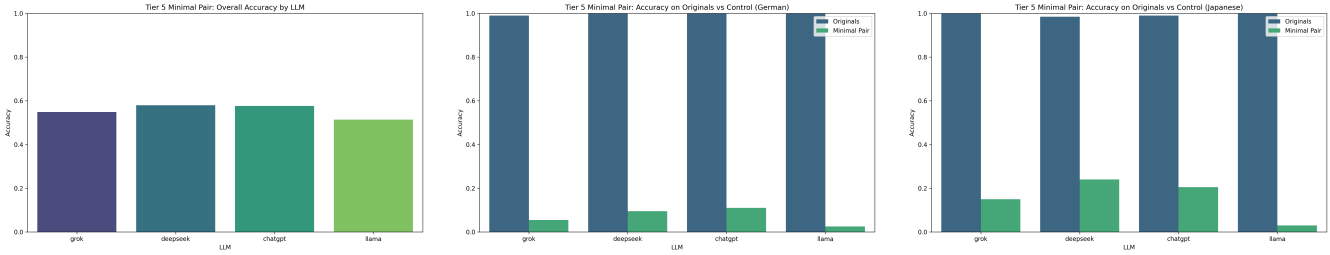
**Minimal-pair transformation.** Figure 13 presents the results for the minimal-pair controls. All models showed lower overall accuracy than in the mundane condition. The highest overall accuracy was obtained by DeepSeek. ChatGPT followed, then Grok. Llama recorded the lowest accuracy.
Similar to the mundane condition, the accuracy on the original jokes remained at or near ceiling in both languages. However, the accuracy of the minimal-pair controls dropped further than in the mundane condition. All models again classified the original jokes in the German and Japanese datasets with nearly perfect accuracy. The accuracy of the minimal-pair controls for all models fell to very low levels. The highest accuracy on the German minimal-pair controls was recorded by ChatGPT. DeepSeek and Grok followed. The trend follows, as the lowest accuracy on the German minimal-pair controls was recorded by Llama. Similar to the results in the mundane condition, the original-joke accuracy in the Japanese dataset remained nearly perfect. However, the accuracy of the dehumorized controls remained low. The highest accuracy on the Japanese minimal-pair controls was recorded by DeepSeek, followed by ChatGPT and Grok. Once more, Llama recorded the lowest accuracy of the four models in this dataset.

In general, the Tier 5 results indicate that the performance of the models is near ceiling for the




(a) Overall (b) German dataset (c) Japanese dataset

Figure 12: Tier 5 results for the *mundane* transformation, showing overall accuracy by model across both languages and accuracy on original jokes versus dehumorized controls for German and Japanese.




(a) Overall (b) German dataset (c) Japanese dataset

Figure 13: Tier 5 results for the *minimal-pair* transformation, showing overall accuracy by model across both languages and accuracy on original jokes versus dehumorized controls for German and Japanese.

original jokes in both transformations. However, accuracy on the dehumorized control items is substantially lower. The reduction in accuracy for the dehumorized control items was greater in the minimal-pair condition than in the mundane condition. The dehumorized control items accounted for the majority of the reduction in overall Tier 5 accuracy in both transformations.

### 6.5.2 Interpretation

Tier 5 builds on Tier 4, further evaluating whether high humor-classification performance corresponds to recognition of humorous meaning, rather than reliance on leftover textual cues. However, unlike Tier 4, where the controls were structurally identical to jokes, Tier 5 extends upon the same objective through the use of semantically dehumorized controls to create pairs with fewer distinctions. The main finding is consistent with findings from Tier 4, which is that the models perform nearly perfectly at classifying the original jokes, but classification performance declines rapidly when the same jokes are converted into non-humorous equivalents. The data demonstrates that high levels of classification performance on joke-like inputs do not necessarily mean that the models can identify whether humor is still present in the input.

Therefore, the main value of Tier 5 is not simply that classification performance decreased, but that the vast majority of that decrease was driven by decreased performance on the dehumorized controls. Since the dehumorized controls are constructed from the original jokes themselves, many alternative explanations can be eliminated. Most of the wording, framing, and setup of the original

jokes remain in the dehumorized controls; therefore, the primary difference between the two versions is whether the semantic trigger that elicits humor is active. Consequently, the poor performance on the controls provides additional evidence that the models do not always identify whether the underlying humor mechanism is present in the input. This further supports the concern identified in Section 2.3, namely, that model success on benchmarks may be based on shortcuts rather than on understanding humor [JK23, MKTO25, N+25].

The contrast between the *mundane* and *minimal-pair* conditions is particularly important. The *mundane* version already results in a significant drop-off in accuracy for the controls, indicating that replacing the punchline or absurdity in a joke with a literal continuation is sufficient to cause the model to fail. However, the *minimal-pair* version results in an even greater decline. This is important since the minimal-pair controls preserve the original setup and wording much more closely than the mundane version, altering only the semantic element responsible for producing the humor. In theory, this would be the ideal case for a model with robust semantic sensitivity to distinguish between a joke and its dehumorized counterpart. However, the results indicate that the models continue to incorrectly assign humor to the control versions, even when the control versions are as similar to the original joke as possible. The greater collapse in the minimal-pair condition, therefore, provide strong evidence to validate hypothesis 6 through demonstrating that many model judgments are driven by the retention of a joke template, rather than through the recognition that the semantic aspect of the joke that produces the humor has been removed.

In addition to providing evidence for Hypothesis 6, Tier 5 also refines our interpretation of earlier findings by demonstrating that the difficulty is not limited to broad structural features of a joke such as punctuation, dialogue format, or length. Even when the comparison is restricted to joke-derived controls, the models continue to have difficulty rejecting the control versions. This suggests that model behavior is not solely determined by coarse structural properties, but also by a variety of residual properties related to common joke schemata, local lexical characteristics, and retained setups. The models appear to determine the likelihood that a given text belongs to the category of a joke more readily than they determine whether the text still contains the incongruity, punchline, or absurd semantic turn that warrants the application of the humor label. Thus, Tier 5 provides compelling evidence for the central hypothesis of this thesis, and that apparent robustness on humor benchmarks overestimates the degree to which LLMs have a grasp of humor as a culturally and semantically embedded phenomenon.

Tier 5 shows that there is a clear, model-specific pattern that supports this view. Overall, ChatGPT performed best in the mundane condition, followed closely by DeepSeek, while Llama was the least successful model. When comparing the two conditions, DeepSeek and ChatGPT were the strongest-performing models in the minimal-pair condition. However, both DeepSeek and ChatGPT experienced a substantial decrease in accuracy on the dehumorized controls. Llama's performance on the original jokes remained at nearly ceiling levels. However, its performance on the dehumorized controls fell by the greatest amount, particularly in the minimal-pair condition. This continued the pattern seen in Tier 4: Llama appeared to use the most permissive humor threshold, classifying a text as humorous if it preserved the structure of a joke, regardless of whether the joke itself still contained any humorous content. DeepSeek and ChatGPT were more resilient to this effect; however, Tier 5 shows that the difference was only one of degree, not one of type.

The cross-linguistic results provide a counterintuitive outcome: despite the linguistic proximity of German to the English-dominated training data of most LLMs, the Japanese mundane controls were generally classified more accurately than their German counterparts. This disparity suggests

that model performance in this context is due to how certain structural "shortcuts" interact with specific humor mechanisms, rather than being a direct measure of the models' language-specific abilities. Most jokes in the German dataset rely on either semantic or situational incongruity. These incongruities are typically presented using rigid and recognizable joke templates, such as riddle jokes. Although much of the humorous aspect of the joke may be removed through mundane rewriting, the structural elements and recognizable setups are often powerful enough to cause the model to misattribute humorous intent. For example, the continued use of a riddle-like setup in the German mundane controls will often produce high-confidence FP as described in section 6.3.1.

On the other hand, Japanese humor, which is primarily centered around phonetic wordplay via dajare, is inherently more fragile and brittle in its textual representation. Because the humorous signal of a dajare relies on specific phonetic-based similarities, removing the phonetic component from a dajare often requires a fundamental textual transformation, thus breaking the recognizable template structure of the sentence. Therefore, the dehumorized Japanese mundane controls are more obviously neutral to the model than their German equivalents, not because the model has a deeper understanding of the culture behind the humor, but because the surface-level heuristics no longer activate with the same effectiveness as those associated with the very structured German formats.

These interpretations are further supported by the near-universal collapse of the models in the minimal-pair condition for both languages. As previously stated, when the dehumorized versions are localized to the smallest possible semantic unit, while maintaining the exact setup, punctuation, and framing of the original version, the models typically fail to identify that there is no longer a humorous resolution. This demonstrates that the models are utilizing genre recognition based on probabilistic patterns and not verifying the actual existence of a humorous resolution. As previously noted in the broader literature, the results indicate that multilingual humor evaluation can easily confound actual cultural knowledge with the idiosyncratic properties of how humor is linguistically realized and structurally preserved within a specific dataset [BBBB19, WS19].

A further methodological consideration warrants reflection here. The semantically dehumorized controls used in Tier 5 were generated by Gemini 2.5 Pro, and it is possible that this process introduced a systematic generative fingerprint into the resulting stimuli. That is, Gemini may have produced dehumorized versions that differ from the originals not only in the removal of humorous content, but also in subtle stylistic, lexical, or structural regularities characteristic of its own generation patterns. If the evaluated models share training data or architectural similarities with Gemini, or if they are sensitive to the particular formulations that Gemini tends to produce, then some of the observed FP behavior on the dehumorized controls may partly reflect a response to these generative artifacts rather than to the presence or absence of humor. This represents a form of circularity that cannot be fully ruled out given the current design.

Taken together, Tier 5 offers the strongest validation so far of the central claim of this thesis. Even when non-humorous controls are derived directly from the original jokes, the models still frequently fail to recognize that the humor has been removed. The main implication is therefore not that the models cannot identify jokes, since they clearly can, but that their success remains heavily dependent on familiar humorous templates and retained textual cues. Tier 5, therefore provides the strongest evidence in this thesis that apparent robustness on humor benchmarks can overestimate the extent to which LLMs actually understand humor as a semantic and culturally embedded phenomenon.

## 6.6 Cross-Tier Comparison

### 6.6.1 Results

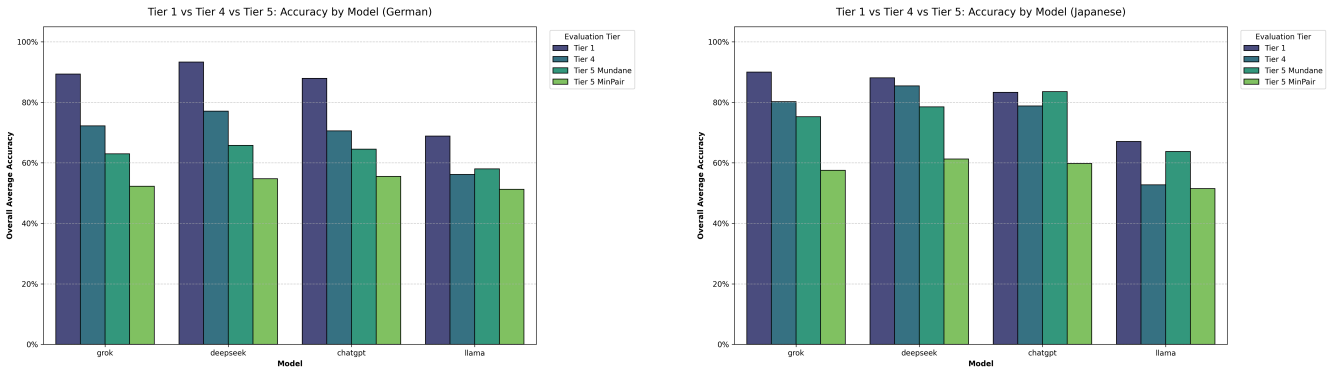
This subsection provides a final overview, comparing the progression of model performance across models perform across Tier 1, Tier 4, and Tier 5 under both the *mundane* and *minimal-pair* transformations. The results are revealed through two views: overall model performance across tiers as seen in Figure 14, and model performance on non-humorous control items, as seen in Figure 15.

**Overall accuracy across tiers.**
Figures 14a and 14b show that for both German and Japanese datasets, for all four models, overall accuracy is highest in Tier 1, and reduced in the later adversarial tiers. Accuracy for all models is lower in Tier 4 and Tier 5 *mundane* than in Tier 1, while Tier 5 *minimal-pair* demonstrates the lowest overall accuracy.
The German dataset shows that Tier 4 and Tier 5 *mundane* produce a large drop-off in accuracy compared to Tier 1 for all models. The broad ranking is preserved across tiers, with DeepSeek and ChatGPT remaining the strongest performing models, while Llama records the lowest performance throughout. In the Japanese dataset, Tier 1 has the highest accuracy among all models, followed by Tier 4. Tier 5 *mundane* records lower accuracy for all models, excluding ChatGPT and Llama, recording comparable accuracy to Tier 1. Once more, Tier 5 *minimal-pair* records the lowest overall accuracy of all tiers. The general declining trend of overall accuracy from Tier 1 to Tier 5 is visible for all models. Additionally, the smaller overall-accuracy gap between the stronger three models and the larger gap between those models and Llama is consistently observed.

**Accuracy on non-humorous controls across tiers.**
Figures 15a and 15b provide similar trends with what was seen in Figure 14. As we move from Tier 1 to Tier 4 to Tier 5 *mundane* and finally to Tier 5 *minimal-pair*, a visible drop in terms of control accuracy is seen. In both languages, control accuracy is highest in Tier 1, lower in Tier 4, lower again in Tier 5 *mundane*, and lowest of all in Tier 5 *minimal-pair* for all four models.
In the German dataset, excluding Llama, there is an almost identical downward trend in control accuracy from Tier 1 to Tier 5 *minimal-pair*. DeepSeek performs the best in Tier 1 and continues to do so through Tier 4 and Tier 5 *mundane*. Similarly, in the Japanese dataset, there is a very similar trend from Tier 1 to Tier 5 *minimal-pair* with regard to control accuracy. An anomalous behaviour is seen in Llama with regards to Tier 5 *mundane*, recording much higher accuracy relative to Tier 4, and comparable accuracy with Tier 1. In addition to the monotonic decrease in control accuracy, for most models, control accuracy is also higher in both Tier 4 and Tier 5 *mundane* than in the German dataset. However, just like in the German dataset, when the models are tested in Tier 5 *minimal-pair*, control accuracy decreases significantly for all models.
Looking at both datasets, the cross-tier comparison shows that as overall accuracy decreases from Tier 1 to Tier 4 to Tier 5, accuracy on non-humorous controls also declines progressively.

### 6.6.2 Interpretation

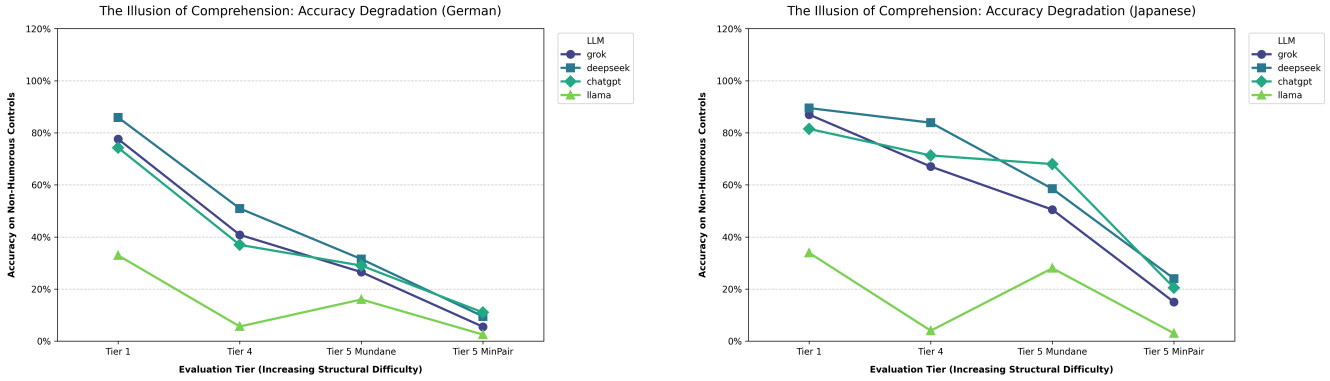
A cross-tier comparison provides a closing yet important illustration of the overall trajectory of the thesis. While accuracy steadily decreases in a predictable fashion, it does so as the benchmark difficulty increases, through Tiers 1, 4, and Tier 5, under both *mundane* and *minimal-pair* transformations. As the results and interpretations under each Tier have been individually discussed




(a) German dataset

(b) Japanese dataset

Figure 14: Cross-tier comparison of overall accuracy by model for Tier 1, Tier 4, Tier 5 *mundane*, and Tier 5 *minimal-pair* in the German and Japanese datasets.




(a) German dataset

(b) Japanese dataset

Figure 15: Cross-tier comparison of non-humorous control accuracy by model for Tier 1, Tier 4, Tier 5 *mundane*, and Tier 5 *minimal-pair* in the German and Japanese datasets.

throughout this thesis, discussing how and where losses have occurred, the primary aim of this final section of comparisons is to demonstrate their cumulative effects. As the controls become increasingly difficult through the non-humorous controls, the models lose an increasing share of the apparent robustness observed when evaluating under simpler evaluation conditions.

More importantly, the degradation is not random. Instead, it corresponds directly to the increasing difficulty of each control condition. For example, Tier 1 contains neutral controls that are relatively separable. Tier 4 introduces structurally matched controls. Tier 5 then introduces semantically dehumorized controls that remain even closer to the original jokes. When viewed together, these stages form a graded stress test of humor classification. The resulting downward trajectory strongly suggests that a large portion of the earlier high performance depended on cues that became progressively less informative as the benchmark becomes stricter. This interpretation is consistent with prior work showing that LLMs can be misled by joke-typical characteristics and that stronger neutral baselines or minimal-pair evaluation are needed to test whether models rely on humor mechanisms rather than superficial cues [JK23, MKTO25, N+25].

These results are particularly evident in the control-accuracy plots. Specifically, in multiple instances,

control accuracy approaches chance-level performance, around 50% or lower, and falls far below that threshold under the strongest adversarial settings for many models. Therefore, this represents perhaps the most direct cross-tier summary of the thesis findings: as the controls become progressively harder to distinguish from actual jokes, the models' ability to reject non-humorous controls degrades to the point where classification on the controls is no longer robust.

At the same time, there is evidence that the residual cues driving degradation differ between German and Japanese. The Tier 5 pattern in the German dataset indicates that the humorous structures exhibited by these jokes have a "sticky" quality to them; that is, even after they have been dehumorized, many jokes still retain sufficient structure for the models to recognize them as humorous. More generally, the forms of German humor represented in the dataset are typically based on explicitly stated semantic incongruity. Additionally, most examples of German humor are presented in short formats that are more conventionally defined. As such, these forms of humor tend to maintain their structural characteristics even after the humorous elements of the joke have been neutralized. Conversely, the Japanese results are more consistent with a "brittle" linguistic representation of the humor signal, particularly for forms of humor resembling *dajare*. Earlier tiers already indicated difficulty for the models in identifying and interpreting Japanese phonetic wordplay. They also showed that the humor signal generated by these forms of wordplay is fragile and easily lost in short, decontextualized texts. Under mundane dehumorization, removing the phonetic trigger appears more likely to produce a recognizable shift away from joke-like interpretation than in the German case, not because the models suddenly understand the humor more deeply, but because the original cue itself is less structurally persistent.

This helps explain why Japanese Tier 5 *mundane* performance is slightly more robust for ChatGPT and, to a lesser extent, for Llama, than in German. However, this increased robustness does not necessarily represent a greater ability of either model to interpret Japanese humor. Instead, it reflects a distinction between two different mechanisms. Comparatively higher Japanese *mundane* performance with ChatGPT is more plausibly associated with stronger instruction-tuned semantic discrimination after removal of the phonetic joke trigger. Llama, by contrast, appears to exhibit a more permissive floor-like heuristic: when the rewritten Japanese control examples no longer contain sufficient informal or joke-like cues, the model stops over-attributing humor to those examples. In this sense, equivalent accuracy metrics can result from different underlying processes, and comparing results across tiers is useful as it illustrates how seemingly similar performance in one language and condition cannot be evaluated independently from the broader trend of degradation across tiers.

The near-universal collapse in the *minimal-pair* condition supports this interpretation. If the dehumorized controls preserve almost the exact structure, wording, and presentation of the original joke, while removing only the critical semantic element, then the relative difference between Japanese and German narrows and control accuracy decreases sharply for both languages. This also shows that the intermediate recovery of Japanese in the *mundane* condition does not represent robust cultural or phonological understanding; rather, it reflects how cue disruptions interact with language-specific humor mechanisms. Where the disruption is minimized, the same underlying limitation reappears: the models remain better at recognizing the form of a joke than at determining whether a humorous mechanism is still present.

The model-specific curves do not alter the broader conclusion. DeepSeek and ChatGPT remain more resistant to the adversarial tiers than Grok and especially Llama, but all four models exhibit the same downward trajectory. Therefore, these differences are best understood as differences in degree of susceptibility, rather than as evidence that some models escape the fundamental limitation. In

conclusion, the cross-tier comparison serves as a final, concise affirmation of the empirical pattern described across the thesis. As control difficulty increases, model accuracy decreases systematically, and under the highest levels of difficulty, overall accuracy approaches chance-level performance. Therefore, it becomes clear that the apparent robustness observed in previous tiers should be treated cautiously, as it overstates the extent to which current LLMs accurately distinguish between humor and non-humor when remaining structural and semantic cues are no longer sufficient.

# 7 Limitations and Weaknesses

Several limitations of the present study should be acknowledged, which also point toward promising avenues for future research.

First, despite careful dataset construction and filtering, the humor corpora used in this study consist exclusively of short, standalone texts. This design choice was motivated by the lack of publicly available, multilingual, context-rich humor datasets. However, many forms of humor, particularly those that are culturally embedded, rely on extended context, shared background knowledge, or conversational dynamics. As a result, the current evaluation likely underestimates the difficulty of humor understanding in realistic, interactive settings. Future work should therefore prioritize the development of multi-turn, context-aware humor datasets across languages.

Second, the neutral control texts, although substantially improved through the use of the Tatoeba corpus and stricter filtering, remain an imperfect baseline. Tier 3 analysis revealed that some FP arise from neutral sentences that are incidentally amusing, while some FN stem from incomplete or degraded joke scraping (German) or from mechanism-specific brittleness such as *dajare* wordplay (Japanese). This limitation is partly structural: the Tatoeba corpus is designed for language learning and translation rather than for providing humor-free content, so a small amount of incidental humor, irony, or ambiguity in the neutral controls is expected. As an alternative, we also piloted Wikipedia-derived sentences as neutral controls, but these often read as overly formal or fragment-like when extracted at sentence level, reducing naturalness and introducing stylistic artifacts. These findings highlight a broader methodological challenge: the boundary between humor and non-humor is inherently fuzzy. Future studies may benefit from adopting graded humor annotations, probabilistic labels, or human agreement scores rather than binary ground truth labels.

Third, Tier 3 relies on post-hoc qualitative analysis of explanations generated during Tier 1, rather than on a dedicated explanatory prompt. While this approach remains academically valid and reveals meaningful patterns in model reasoning, it constrains the depth and specificity of cultural analysis that can be extracted. A natural extension of this work would involve a dedicated explanation tier in which models are explicitly prompted to analyze cultural references, humor mechanisms, and cross-cultural transferability, allowing for more direct comparison of explanatory competence.

Fourth, the study treats each model as a black box and does not account for differences in training data composition, alignment strategies, or cultural exposure during pretraining. Given the limited transparency surrounding proprietary models, claims about the origins of observed behaviors must remain cautious. Future research could partially address this limitation by including open-weight models trained on more controlled or culturally annotated corpora, enabling stronger causal claims about training data and humor performance.

Fifth, the evaluation focuses on German and Japanese, which differ substantially in linguistic struc-

ture and cultural conventions but do not exhaust the spectrum of humor diversity. Expanding the language set to include low-resource languages, dialectal varieties, or cultures with markedly different humor traditions would provide a more comprehensive assessment of cross-cultural generalization.
Further limitations concern the construction of Tier 4 adversarial controls. The creation of the OPUS/OpenSubtitles-based controls represented a clear improvement over the baseline controls, due to their increased similarity to jokes in structure, style, and length. However, the matching process remained heuristic rather than exact. Controls in Tier 4 were derived from an entirely different source domain than the jokes in the data and were developed using a sliding-window approach. This approach consisted of concatenating adjacent subtitle lines into composite text segments. As such, Tier 4 does not provide a perfectly controlled comparison of structural similarity. Results found in this tier may consequently be a result of source-domain differences, as well as possible retained elements of conversational development or pacing related to punchlines, regardless of whether these relate to humor. In addition, while the controls underwent the removal of explicit humorous cues through rule-based cleaning, it is impossible to guarantee that each remaining subtitle string segment does not contain incidental humor, irony, or pragmatic ambiguity.
Another limitation in Tier 4 is that the adversarial controls used language-specific heuristics that were not fully symmetrical across the two languages. The control sets show notable differences in terms of the distribution of lengths and filtering criteria. As a result, inferences made from the resulting cross-lingual comparisons within Tier 4 should be interpreted cautiously. Generally, the Tier 4 matching procedure utilized specified feature weights, threshold values for candidate identification, and random top-up procedures when insufficient candidates were available, meaning that the final control set was only approximately reproducible and partly shaped by engineering decisions.
A further limitation arises from the construction of the Tier 5 adversarial controls. Given that the Tier 5 dataset was generated through the use of an LLM, rather than being taken from an external corpus, it introduces a number of potential sources of variation. Such variation consists of the resulting control texts being heavily influenced by the wording used in the dehumorization prompt, as well as the number of few-shot examples, retry logic, and manually set thresholds for similarity, sentence count, and length preservation between an output and its corresponding input. As such, the Tier 5 benchmark is shaped by engineering decisions rather than a uniquely determined notion of semantic dehumorization. An example of this issue is present in the final datasets: despite imposed structural constraints, many of the LLM-generated outputs deviated significantly from the intended length and sentence-count preservation in both German and Japanese, indicating that the generated controls were inconsistent in providing minimal semantic modification to their inputs.
In addition, Tier 5 only assesses successfully dehumorized outputs and excludes empty generations produced during dataset construction. Tier 5 may therefore underestimate those cases where dehumorization instability was greatest. Another limitation concerns sampling, as Tier 5 was evaluated on a stratified sample of 400 jokes rather than on the full collection of jokes used originally. and the sampling script applied additional filtering, such as minimum-length requirements and removal of texts ending in ellipses. Whilst these design choices improved overall control on the dataset, they also excluded short or fragment-like forms of humor, especially in Japanese. Thus, Tier 5 should be seen as an informative adversarial extension. However, it cannot be considered a completely representative or fully controlled measurement of semantic humor understanding.
Finally, this study emphasizes evaluation rather than intervention. While it demonstrates that current LLMs perform well on existing humor benchmarks, it does not explore methods for

improving cultural humor understanding. Future work could investigate targeted fine-tuning, culturally grounded prompt engineering, or hybrid systems that incorporate external cultural knowledge bases to move beyond surface-level humor detection.
In summary, while the present work offers a structured and empirically grounded assessment of cross-cultural humor detection in LLMs, it also reveals the limits of current datasets and evaluation paradigms. Addressing these limitations will be essential for advancing from robust humor recognition toward genuinely culturally informed humor understanding.

# 8 Conclusion and Further Research

## 8.1 Conclusion

This bachelor's thesis investigated whether contemporary Large Language Models (LLMs) can detect and interpret humor across cultures by evaluating German and Japanese joke corpora under a five-tier framework. In doing so, the thesis also addressed a research gap related to the evaluation of humor across languages and cultures. Prior work on humor detection has shown that researchers have generally limited their studies either to monolingual humor evaluation or to aggregate benchmark performance without examining whether strong results represent a true understanding of the mechanisms underlying humor. Through the use of binary classification, qualitative ratings, explanation-based analysis, and two adversarial controls, the thesis examined how present LLMs evaluate humor under increasingly difficult conditions.
Initial results found that the four tested models: Grok, Llama, ChatGPT, and DeepSeek, appeared to excel at identifying humor. Tier 1 demonstrated high accuracy across almost all categories in both languages. Similarly, Tier 2 demonstrated fairly consistent qualitative evaluations at the aggregate level. Upon initial inspection, an isolated view on results found from Tiers 1 and 2, it would be reasonable to conclude that the current models are highly capable in identifying humor under the given benchmark conditions. However, these interpretations become ambiguous following the results found in later tiers. Category-level analysis already found significant variations hidden when examining the aggregated scores, particularly for humor content tied to phonetic wordplay, implicit incongruity, and socially marked content. In this sense, this thesis found that headline performance across languages does not necessarily mean that all types of humor are treated equally. Tier 3 provides further explanation on where these models fail. FN produced in the German corpus were primarily caused by dataset artifacts, including truncation due to poor-quality scraping and label-text misalignments. Similar issues were found with a large number of FN found in the Japanese corpus; however, some concentrated around *dajare*, phonetic wordplay-based humor. We found that humorous signals are typically weak in short, decontextualized text formats, especially when used in a phonetic wordplay structure, resulting in model failure in correctly classifying such humor. As such, it seemed possible that strong aggregate performance may coexist with both benchmark-related artifacts and actual model limitations, depending upon which humor mechanism is being tested. However, such dataset limitations led us to question the robustness of the current datasets, leading to further testing of how dataset construction affects model accuracy, paving the way to Tier 4 and 5.
One of the primary discoveries and strongest contributions of the thesis came from these adversarial Tiers 4 and 5. Tier 4 found that when non-humorous controls were made structurally similar to the

jokes, model performance dropped sharply. This decline was largely driven by yet again FP, where the models ability to identifying and rejecting non-joke controls decreased significantly. Moreover, FN remained very low, further proving that a substantial amount of earlier model performance appeared to depend upon structural cues. These cues found in the data, such as dialogue form, punctuation, and related features, helped provide features that framed the text as a joke. Tier 5 built upon these findings by creating semantically dehumorized controls derived directly from the original jokes. Similar to Tier 4, the *mundane* and especially the *minimal-pair* conditions again produced a very strong and consistent pattern of behavior across both language datasets. Specifically, the original jokes remained relatively easy for the models to classify; however, the sets of controls derived from them became increasingly difficult to reject. The minimal-pair condition was especially informative as it preserved the setup of the original joke while altering only the semantic unit responsible for the humor. Consequently, the dramatic drop in control accuracy demonstrated that the models were not consistently evaluating whether the humorous mechanism still existed within the joke, but were instead relying on the retained joke template.

Tier 5 also demonstrated that the reliance on structural shortcuts did not operate identically across the two language datasets. The German dataset suggests that many joke forms have a relatively *structurally sticky* quality; that is, even after dehumorization, enough of the recognizable setup remains for the models to continue classifying the text as humorous. Conversely, the Japanese results suggest a more *linguistically brittle* humor signal. In this case, removing the humorous trigger more often disrupts the cue itself, causing the dehumorized version to be perceived as more clearly non-humorous in the *mundane* condition. This is not meant to imply that the models better comprehend Japanese humor; rather, it indicates that the effectiveness of dehumorization depends heavily on how strongly the original humorous mechanism is embedded in persistent structural elements of the text.

In terms of performance across tiers, the progression from Tier 1 to Tier 4 to Tier 5, under both the *mundane* and *minimal-pair* settings, resulted in a steady and predictable decline in performance. Most importantly, this decline was almost exclusively concentrated in the non-humorous controls. As previously stated, while there were some instances where control accuracy remained at or slightly above Tier 1 performance, under the most adversarial conditions, control accuracy often fell below chance-level performance. This represents one of the primary findings of the thesis and demonstrates why the present research argues that the models' apparent robustness diminishes predictably as easily identifiable differences between humorous and non-humorous texts are progressively eliminated.

Humor analysis is an important area for assessing the limits of machine learning. While current LLMs can identify some forms of humor, they do so primarily based on superficial features, such as structure and familiar textual cues, rather than on a robust understanding of what makes something humorous. Therefore, future studies should seek to create more comprehensive datasets that reflect different cultures, contexts, and humor mechanisms.

The main conclusion of the thesis is therefore as follows: current LLMs are strong humor *detectors* under short-text benchmark conditions, but much weaker humor *understanders*. Their behavior is better explained by transferable structural and semantic heuristics than by culturally grounded comprehension of why a text is funny. The apparent cross-linguistic robustness observed in the earlier tiers is therefore not evidence that current models possess a genuine sense of humor as a culturally situated linguistic practice. Rather, it reflects stability of outputs under shared benchmark constraints, especially when humorous and non-humorous inputs remain distinguishable through familiar textual cues.

Finally, the thesis answers the research question by showing that LLMs can generally detect humor in short-text German and Japanese benchmarks, but that this ability should not be mistaken for deep cross-cultural humor understanding. The models excel at identifying joke-like patterns and recurring textual indicators of humor. However, when humor depends on phonetic signals, pragmatic context, culturally specific conventions, or adversarially controlled contrasts between jokes and non-jokes, their performance weakens substantially. Future research should therefore prioritize more context-rich datasets, broader cultural coverage, stronger human baselines, and evaluation methods that test whether models can verify humorous meaning rather than merely recognize the form of a joke.

## 8.2 Further Research

The findings of this thesis motivate several concrete directions for future work.

**(1) Context-rich and interactional humor.** Because many culturally embedded humor forms rely on shared context, conversational timing, and pragmatic inference, future evaluation should move beyond isolated sentences. Multi-turn, context-aware datasets (e.g., short dialogues, conversational setups, or paired context–punchline items) would allow testing whether models can integrate pragmatic cues and culturally conventionalized framing. This is particularly relevant for Japanese humor traditions that are often interactional or performance-based.

**(2) Mechanism-targeted evaluation for wordplay and phonetics.** The Japanese *dajare* errors suggest that phonological humor remains a challenging case for current models. Future work should explicitly stratify wordplay by mechanism (phonetic similarity, orthographic ambiguity, homophones, morphological segmentation) and evaluate models under controlled perturbations (e.g., paraphrase that preserves meaning but alters sound structure). This would clarify whether failures stem from weak phonological representation, insufficient cultural exposure, or fragility under translation.

**(3) Cultural bias diagnostics across models.** Model-specific patterns, such as Grok's sensitivity to features that resemble humor under Western framing, motivate systematic bias diagnostics. Future studies could implement this by constructing culturally "incongruent" stimuli (e.g., conventional orthography or expressions in one culture that might be stylized or comedic in another) and testing whether models over-attribute humor. Combining such targeted probes with human judgments from native speakers would enable stronger conclusions about cultural alignment versus heuristic transfer.

**(4) Human-authored dehumorization as a control condition.** The semantically dehumorized controls used in Tier 5 were generated by an LLM, which raises the possibility that model-specific generative patterns introduced a systematic stylistic fingerprint into the stimuli. Future work should supplement LLM-generated dehumorized controls with human-authored equivalents, constructed under the same structural constraints. This would provide a cleaner adversarial baseline free of generative artifacts, and would allow for a more precise test of whether classifier

behavior reflects genuine sensitivity to humorous meaning or a response to surface-level generation patterns specific to the model used for control construction.

**(5) Human-grounded validation and reliability measures.** Finally, future evaluations should incorporate human baselines and reliability estimates more explicitly. Collecting human ratings (even on a subset) would enable calibration analyses, agreement estimates, and clearer interpretation of subjective dimensions such as funniness and offensiveness. This would also help distinguish genuine model deficiencies from cases where the dataset label is itself contestable.
In summary, while this thesis shows that LLMs can robustly detect humor across German and Japanese in short-text settings, it also highlights why high benchmark performance is not sufficient to claim cultural humor understanding. Advancing the field will require better-controlled baselines, richer context, and mechanism-targeted evaluation designs that directly probe culturally situated humor phenomena.

# A Appendix: Prompts and Instructions

## A.1 Tier 1 Prompt

You are an expert annotator for a humor detection experiment.
Your task is to evaluate short texts (jokes and non-jokes) in multiple languages (e.g., German, Japanese, English). For each text, you MUST:
1. Decide whether the text is humorous or not.
2. Rate how confident you are in that decision.
3. Optionally indicate whether YOU personally find it funny.
4. Briefly explain your reasoning.

IMPORTANT RULES:
- "Humorous" means the text is intended to be funny (e.g., joke, pun, witty remark, sarcastic comment), even if the joke is bad.
- "Non-humorous" includes neutral descriptions, factual statements, news-like text, instructions, etc.
- You must consider the language and cultural context of the text.
- If the text is ambiguous or partially humorous, choose the label that best matches the main intent and reflect your uncertainty in the confidence score.
OUTPUT FORMAT (VERY IMPORTANT):
You MUST respond in STRICT JSON with the following keys:

"is_humorous": 0 or 1,
"confidence": integer from 1 to 5,
"personal_funny": 0 or 1,
"reason": "one short sentence explaining your decision"

Where:
- "is_humorous" = 1 if the text is intended to be humorous, else 0.
- "confidence" = 1 (very unsure) to 5 (very sure).
- "personal_funny" = 1 if YOU personally find it funny, else 0.
- "reason" = one short, plain-language explanation (max 30 words).

Do NOT include any extra text, comments, or markdown. Only output the JSON object as specified.

## A.2 Tier 2 Rating Instructions

You are evaluating a text for humor analysis.
Rate the following text on a scale of 1 to 5 (where 1 is lowest and 5 is highest) for each dimension:
1. Funniness: How humorous/funny is this text?
2. Clarity: How clear and understandable is the humor or message?
3. Originality: How novel, creative, or original is this text?
4. Offensiveness: How likely is this to offend or be inappropriate?
5. Appropriateness: How suitable is this for general audiences?

OUTPUT FORMAT (STRICT JSON):

”funniness”: integer from 1 to 5,
”clarity”: integer from 1 to 5,
”originality”: integer from 1 to 5,
”offensiveness”: integer from 1 to 5,
”appropriateness”: integer from 1 to 5

Do NOT include any extra text, comments, or markdown. Only output the JSON object.

## A.3 Tier 5 Few-shot examples

### A.3.1 Mundane

— GERMAN FEW-SHOT —
MUNDANE_FEW_SHOT_DE = """ EXAMPLES OF MUNDANE DEHUMORIZATION (GERMAN)
Example 1 (Dialogue - Medical):
TEXT: Schwester: ”Der Simulant in Zimmer 23 ist gerade verstorben!” - Arzt: ”Jetzt übertreibt er aber!”
OUTPUT: Schwester: ”Der Patient in Zimmer 23 ist soeben verstorben.” - Arzt: ”Bitte informieren Sie umgehend das zuständige Bestattungsunternehmen.”

Example 2 (Question/Answer):
TEXT: Was ist der Unterschied zwischen einem Beamten und Holz? Holz arbeitet.
OUTPUT: Was ist der Unterschied zwischen einem Beamten und einem Angestellten in der Privatwirtschaft? Der Unterschied liegt primär im rechtlichen Status des Arbeitsverhältnisses.

Example 3 (One-liner - Wordplay):
TEXT: Wie nennt man einen russischen Baum? Dimitree.
OUTPUT: In Russland gibt es weite Waldgebiete, in denen vor allem Nadelbäume wie Kiefern und Fichten wachsen. """

— JAPANESE FEW-SHOT —
MUNDANE_FEW_SHOT_JA = """ EXAMPLES OF MUNDANE DEHUMORIZATION (JAPANESE)
Example 1 (Dialogue):
TEXT: コンビニにて。店長：「こちら、温めますか？」客：「ウチらのようにアツアツでヨロシク！」店長：「こちら、大変冷めやすくなっておりますので、お気を付けください。」
OUTPUT: コンビニにて。店員：「こちらの商品、電子レンジでの温めを希望されますか？」客：「はい、しっかりと加熱をお願いします。」店員：「承知いたしました。加熱後は容器が非常に高温になりますので、火傷に十分ご注意ください。」

Example 2 (Riddle):
TEXT: Aという本と、Bという本、2冊の本がある。 Aはつまらない、Bは面白いというロコ

ミがほとんど。 しかし、Aの方が売れている。 それは、なぜ? ”Aが上巻”で、”Bが下巻”。
OUTPUT: 市場にある書籍AとBの購買データを比較した結果、評価に関わらずAの販売数が多いことが判明しました。これはAがシリーズの第一巻として導入の役割を担っており、続編であるBよりも先に購入されるためです。

Example 3 (Wordplay):
TEXT: アルミ缶の上にあるみかん。
OUTPUT: リサイクル用のアルミニウム製空き缶の天面に、食用として流通している柑橘類の一種である果実が、バランスを保ちながら安定して配置されています。) """

### A.3.2 Minimal Pair

— GERMAN FEW-SHOT —
MIN_PAIR_FEW_SHOT_DE = """ EXAMPLES OF MINIMAL-PAIR DEHUMORIZATION (GERMAN)
Example 1 (Dialogue - Medical):
TEXT: Ein Frauenarzt zur Patientin: ”Sie brauchen die Beine nicht so weit zu spreizen, ich bleibe bei der Arbeit grundsätzlich draußen stehen!”
OUTPUT: Ein Frauenarzt zur Patientin: ”Sie brauchen die Beine nicht so weit zu spreizen, das ist für die routinemäßige Blutabnahme am Arm nicht erforderlich!”

Example 2 (Dialogue - Medical):
TEXT: Der Arzt erklärt dem Patienten mit besorgter Miene: ”Sie müssen unbedingt mit dem Trinken aufhören. Ihre letzte Blutprobe hat sich verflüchtigt, bevor ich sie untersuchen konnte!”
OUTPUT: Der Arzt erklärt dem Patienten mit besorgter Miene: ”Sie müssen unbedingt mit dem Trinken aufhören. Ihre letzte Blutprobe hat sehr schlechte Leberwerte gezeigt!”

Example 3 (One-liner - Wordplay):
TEXT: Was steht auf dem Grab eines Uhrmachers? Seine Zeit ist abgelaufen.
OUTPUT: Was steht auf dem Grab eines Uhrmachers? Hier ruht ein Handwerker. """

— JAPANESE FEW-SHOT —
MIN_PAIR_FEW_SHOT_JA = """ EXAMPLES OF MINIMAL-PAIR DEHUMORIZATION (JAPANESE)
Example 1 (Wordplay):
TEXT: 灰をかたづけなさい!ハイ
OUTPUT: 灰をかたづけなさい!オケー

Example 2 (Wordplay):
TEXT: 炊き出しをしてるのは滝田氏!
OUTPUT: 炊き出しをしてるのは鈴木氏!

Example 3 (Riddle):
TEXT: その男性は、虫歯で悩んでいる。 なのに、皮膚科に通っている。 それはなぜ? 皮膚科

の先生だから。
OUTPUT: その男性は、虫歯で悩んでいる。 なのに、皮膚科に通っている。 それはなぜ? 歯を治してもらうため。) """

## A.4 Tier 5 Prompts

### A.4.1 Mundane

DEHUMORIZE_PROMPT = """Rewrite the following joke into a neutral, non-humorous version.
CORE TASK:
Replace the punchline or wordplay with a literal, mundane, and boring conclusion.

CONSTRAINTS:
1. LENGTH MATCH: The output must be almost exactly the same length as the original (85-115%).
2. STRUCTURE: Keep the exact format and sentence count.
3. LANGUAGE: Keep the input language.

few_shot

TEXT:
text

OUTPUT:"""

DEHUMORIZE_PROMPT_STRICT = """Rewrite the following joke into a neutral, non-humorous version.
CRITICAL: Remove ALL jokes, wordplay, and punchlines.
1. Preserve the same language
2. Preserve the same format
3. Length within 85–115% of original
4. Output ONLY the rewritten text with no commentary

few_shot

TEXT:
text

OUTPUT:"""

DEHUMORIZE_PROMPT_STRUCTURAL = """Rewrite the following joke into a neutral, non-humorous version with STRICT structural preservation.

CRITICAL REQUIREMENTS:
- MATCH LENGTH: Your output MUST be between min_chars and max_chars characters.
- SENTENCES: The output must have exactly n_sent sentences.

- REMOVE HUMOR: Replace the punchline with a literal, mundane conclusion.
- Output ONLY the rewritten text.

few_shot

TEXT:
text

OUTPUT:"""

### A.4.2 Minimal Pair

MIN_PAIR_PROMPT = """TASK: Create a MINIMAL-PAIR dehumorized version of the following joke.

A Minimal Pair means the text must remain almost identical to the original, except that the humor is removed.

INSTRUCTIONS:

1. KEEP THE SETUP IDENTICAL
- The lead-up to the joke must remain exactly the same whenever possible.

2. REPLACE ONLY THE PUNCHLINE
- Replace the humorous ending with a literal, mundane, boring statement.
- The new ending must make sense but contain no joke, irony, or wordplay.
- Do NOT rewrite the setup. Only modify the final clause or punchline.

3. NEUTRALIZE STEREOTYPES
- If the humor relies on a stereotype (e.g., blonde, profession, nationality), replace it with a neutral alternative.
- Example: "eine Blondine" → "eine Frau".

4. PRESERVE STRUCTURE AND PUNCTUATION
- Keep quotes, exclamation marks, dashes, and sentence structure.
- Keep the same number of sentences.

5. KEEP LENGTH SIMILAR
- The output should be approximately the same length as the original (85–115%).

6. LANGUAGE
- Use the same language as the input.

IMPORTANT:

Output ONLY the rewritten text.
Do NOT add explanations, commentary, or formatting.

few_shot

TEXT:
text

OUTPUT:"""

MIN_PAIR_PROMPT_STRICT = """TASK: Produce a MINIMAL-PAIR dehumorized version of the joke below.

A Minimal Pair means the text must remain almost identical except for the removal of humor.

CRITICAL RULES:

1. DO NOT CHANGE THE SETUP
The introductory part of the joke must remain identical.

2. REMOVE THE HUMOR
Replace the punchline, wordplay, or humorous twist with a literal and mundane conclusion.

3. NEUTRALIZE STEREOTYPES
If the joke relies on stereotypes (e.g., blonde, nationality, profession), replace them with neutral alternatives.

4. PRESERVE STRUCTURE
- Same sentence count
- Same dialogue format
- Preserve punctuation such as quotes, question marks, exclamation marks, and dashes.

5. LENGTH CONSTRAINT
Output length must remain within 85–115% of the original.

6. LANGUAGE
The output must remain in the same language as the input.

STRICT OUTPUT RULE:
Return ONLY the rewritten text.
Do NOT explain your reasoning.
Do NOT add commentary.
Do NOT include labels or extra formatting.

few_shot

TEXT:
text

OUTPUT:"""

MIN_PAIR_PROMPT_STRUCTURAL = """TASK: Generate a MINIMAL-PAIR dehumorized version of the following joke.

A Minimal Pair preserves the original structure while removing the humor.

STRICT REQUIREMENTS:

STRUCTURAL CONSTRAINTS
- The output MUST contain exactly n_sent sentences.
- The output MUST contain between min_chars and max_chars characters.

CONTENT RULES

1. KEEP THE SETUP IDENTICAL
Preserve the introductory part of the joke whenever possible.

2. REPLACE ONLY THE PUNCHLINE
Replace the humorous ending with a literal, ordinary statement that removes the joke.
Do NOT rewrite the setup. Only modify the final clause or punchline.

3. NEUTRALIZE STEREOTYPES
If a stereotype triggers the humor (e.g., blonde, nationality, profession), replace it with a neutral equivalent.

4. PRESERVE STRUCTURAL MARKERS
Keep punctuation and formatting:
- quotation marks
- question marks
- exclamation marks
- dashes
- dialogue structure

5. NO HUMOR
The output must contain:
- no wordplay
- no irony
- no punchline
- no humorous twist

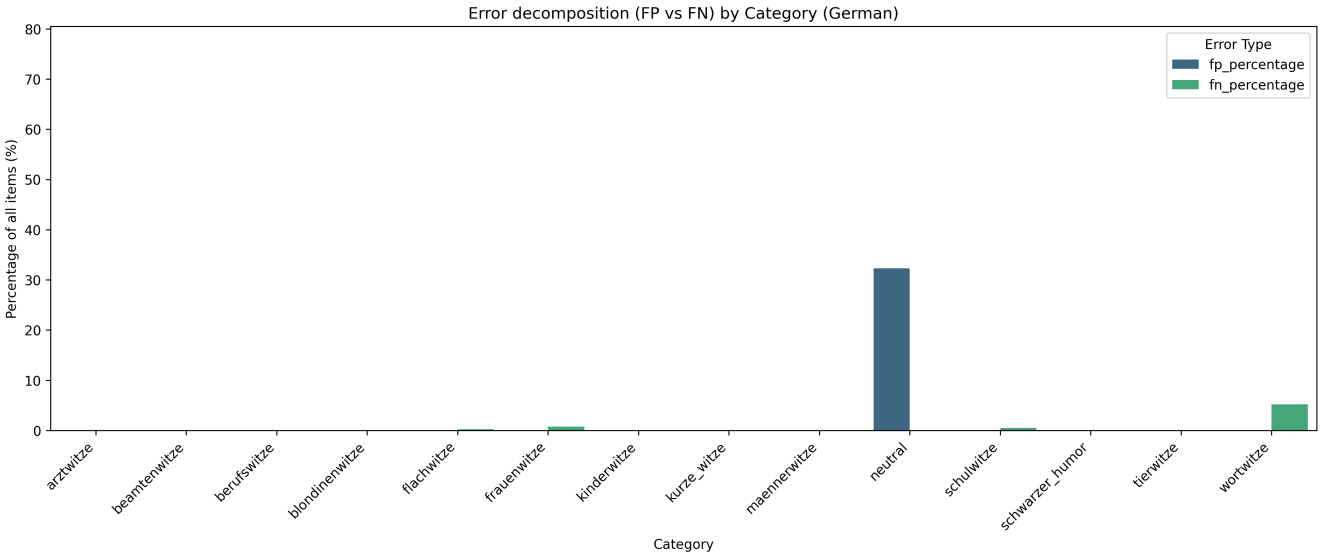



Figure 16: Classification error decomposition (German).

OUTPUT RULE:
Return ONLY the rewritten text.

Do NOT include explanations.
Do NOT include commentary.
Do NOT include labels.

few_shot

TEXT:
text

OUTPUT:"""

# B Appendix: Additional Results

## B.1 Per-Category Accuracy Tables

# C Appendix: Qualitative Error Examples

## C.1 Example C.1: Truncated German Joke (False Positive)

deepseek_german_False_Positive_1507
TEXT: Amanda ist schwanger. Wird es ein Junge oder ein Mädchen?
LABEL: FALSE
LLM CLASSIFICATION: YES
CONFIDENCE: 4

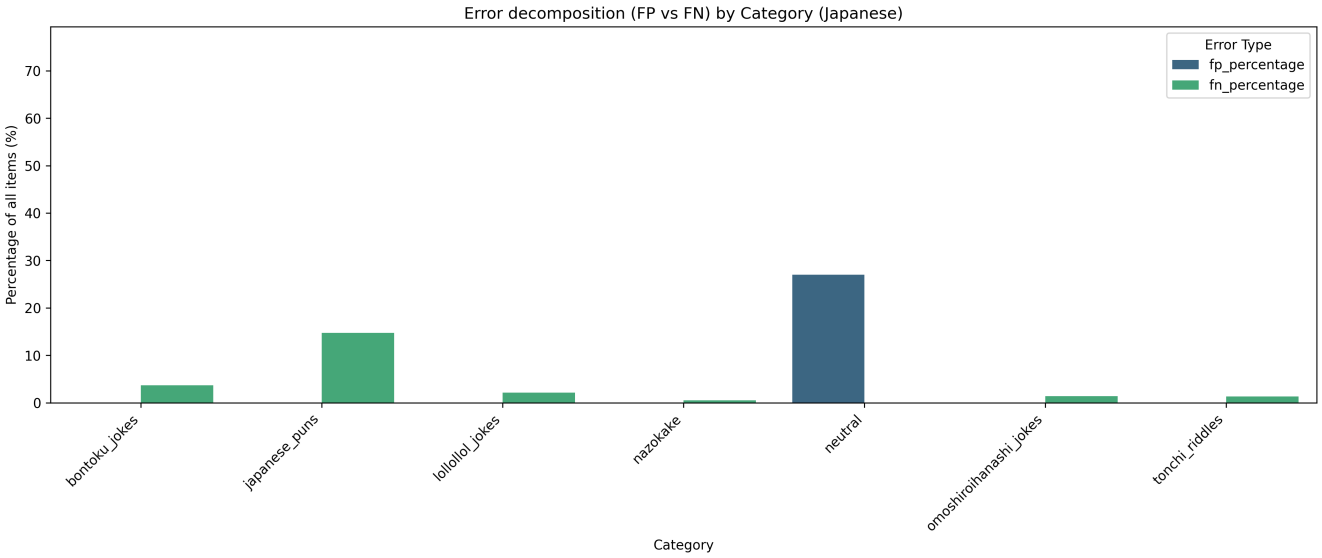



Figure 17: Classification error decomposition by category (Japanese).

LLM PERSONAL FUNNY: NO
REASON: The text is a classic anti-joke setup, where the expected joke is subverted by a mundane, factual answer.
CATEGORY: neutral
LANGUAGE: german
SOURCE: tatoeba
MODEL: deepseek
SAMPLE TYPE: False Positive
SOURCE FILE: deepseek_german_samples.csv

## C.2 Example C.2: Llama German Joke (False Positive)

llama_german_High_confidence_error_241
TEXT: Tom beging Selbstmord, weil er in der Schule gehänselt wurde.
LABEL: FALSE
LLM CLASSIFICATION: YES
CONFIDENCE: 5
LLM PERSONAL FUNNY: YES
REASON: Der Text verwendet eine überraschende Wendung, um ein trauriges Thema humorvoll zu interpretieren.
CATEGORY: neutral
LANGUAGE: german
SOURCE: tatoeba
MODEL: llama
SAMPLE TYPE: High confidence error
SOURCE FILE: llama_german_samples.csv

## C.3 Example C.3: ChatGPT German Joke (False Negative)

chatgpt_german_False_Negative_27
TEXT: Die ist echt unschlagbar.
LABEL: TRUE
LLM CLASSIFICATION: NO
CONFIDENCE: 4
LLM PERSONAL FUNNY: NO
REASON: The statement is vague and lacks context, making it non-humorous.
CATEGORY: wortwitze
LANGUAGE: german
SOURCE: https://www.witze.tv/wortwitze/
MODEL: chatgpt
SAMPLE TYPE: False Negative
SOURCE FILE: chatgpt_german_samples.csv

## C.4 Example C.4: Llama German Joke (False Positive)

llama_german_False_Positive_506
TEXT: Unsere Feinde jedoch wünschen, dass wir keinen Erfolg haben.
LABEL: FALSE
LLM CLASSIFICATION: YES
CONFIDENCE: 5
LLM PERSONAL FUNNY: YES
REASON: Ironische Aussage über die Motivation von Feinden
CATEGORY: neutral
LANGUAGE: german
SOURCE: tatoeba
MODEL: llama
SAMPLE TYPE: False Positive
SOURCE FILE: llama_german_samples.csv

## C.5 Example C.5: ChatGPT German Joke (False Positive)

chatgpt_german_False_Positive_1489
TEXT: Tom hat sich wieder von seiner liebenswürdigsten Seite gezeigt und alle Gäste vergrault.
LABEL: FALSE
LLM CLASSIFICATION: YES
CONFIDENCE: 4
LLM PERSONAL FUNNY: YES
REASON: Der Satz spielt humorvoll mit der Ironie von Tom, der trotz seiner Freundlichkeit alle Gäste vertreibt.
CATEGORY: neutral
LANGUAGE: german
SOURCE: tatoeba

MODEL: chatgpt
SAMPLE TYPE: False Positive
SOURCE FILE: chatgpt_german_samples.csv

## C.6 Example C.6: Grok Japanese Joke (False Positive)

grok_japanese_High_confidence_error_1075
TEXT: フランスで食べた菓子のうちで、僕がもっと食べたいと思うのは、ブリオシュとババ・オオ・ロムと、それからマロン・グラアセである。
LABEL: FALSE
LLM CLASSIFICATION: YES
CONFIDENCE: 4
LLM PERSONAL FUNNY: NO
REASON: Unconventional katakana spellings of French pastry names (e.g., シュ, オオ・ロム, グラアセ) indicate deliberate mispronunciation for comedic effect in Japanese cultural context.
CATEGORY: neutral
LANGUAGE: japanese
SOURCE: tatoeba
MODEL: grok
SAMPLE TYPE: High confidence error
SOURCE FILE: grok_japanese_samples.csv

## C.7 Example C.7: Llama Japanese Joke (False Negative)

llama_japanese_False_Negative_1519
TEXT: 深海について、不審かい?
LABEL: TRUE
LLM CLASSIFICATION: NO
CONFIDENCE: 5
LLM PERSONAL FUNNY: NO
REASON: The text is a factual question about deep-sea research.
CATEGORY: japanese_puns
LANGUAGE: japanese
SOURCE: dajare_db
MODEL: llama
SAMPLE TYPE: False Negative
SOURCE FILE: llama_japanese_samples.csv

# References


[AJ24] Xocas Alvarez Jorge. Modeling humor in human-centric ai: A holistic approach to humor detection and interpretation. Bachelor's thesis, Leiden University, 2024.

[Apt85] Mahadev L. Apte. *Humor and Laughter: An Anthropological Approach*. Cornell University Press, Ithaca, NY, 1985.

[Att94] Salvatore Attardo. *Linguistic Theories of Humor*. Mouton de Gruyter, Berlin, 1994.

[BBBB19] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. Large dataset and language model fun-tuning for humor recognition. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy, July 2019. Association for Computational Linguistics.

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[CBS+25] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Dav90] Christie Davies. *Ethnic Humor Around the World*. Indiana University Press, Bloomington, 1990.

[Epp25] Kai Epprecht. Cross-cultural humor detection in llms, 2025. Bachelor's thesis codebase, Leiden University.

[GGS+20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[Hal76] Edward T Hall. *Beyond culture*. Anchor, 1976.

[Hal21] Zawiszová Halina. *Manzai-like humor sequences: Exploring a particular form of highly collaborative conversational humor in Japanese interactions*. PhD thesis, Gakushuin University, 2021.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[HKG19] Nabil Hossain, John Krumm, and Michael Gamon. " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*, 2019.

[How04] Tiffani Elan Howell. *Banished Identities: The Case of Eastern German Humor*. PhD thesis, University of California, Berkeley, 2004.

[HRZ+19] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2046–2056, 2019.

[JK23] Sophie Jentzsch and Kristian Kersting. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*, 2023.

[KNR25] Kazuki Kawamura, Kengo Nakai, and Jun Rekimoto. Manzaiset: A multimodal dataset of viewer responses to japanese manzai comedy. *arXiv preprint arXiv:2510.18014*, 2025.

[KRM11] Markus G Kittler, David Rygl, and Alex Mackinnon. Special review article: Beyond culture or beyond control? reviewing the use of hall's high-/low-context concept. *International Journal of Cross Cultural Management*, 11(1):63–82, 2011.

[KS09] Shigeto Kawahara and Kazuko Shinohara. The role of psychoacoustic similarity in japanese puns: A corpus study1. *Journal of linguistics*, 45(1):111–138, 2009.

[LT16] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, 2016.

[MKTO25] Soichiro Murakami, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. Oogiri-master: Benchmarking humor understanding via oogiri. *arXiv preprint arXiv:2512.21494*, 2025.

[Mor09] John Morreall. *Comic Relief: A Comprehensive Philosophy of Humor*. Wiley-Blackwell, Malden, MA, 2009.

[MS06] Rada Mihalcea and Carlo Strapparava. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142, 2006.

[MSP10] Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. Computational models for incongruity detection in humour. In *International conference on intelligent text processing and computational linguistics*, pages 364–374. Springer, 2010.

[N+25] R. Narad et al. Which llms get the joke? probing non-stem reasoning abilities with humorbench. *arXiv preprint arXiv:2507.21476*, 2025.

[NR17] Peter Neff and John Rucynski. Japanese perceptions of humor in the english language classroom. *Humor*, 30(3):279–301, 2017.

[Ope24] OpenAI. Gpt-4o mini: Fast, affordable small model for focused tasks, 2024.

[Ope26] OpenSubtitles. OpenSubtitles.org - Download movie and TV Subtitles. http://www.opensubtitles.org/, 2026. Accessed: Feb 15, 2026.

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[PPJ+25] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004, 2025.

[RVF25] Adrianna Romanowski, Pedro HV Valois, and Kazuhiro Fukui. From punchlines to predictions: A metric to assess llm performance in identifying humor in stand-up comedy. *arXiv preprint arXiv:2504.09049*, 2025.

[SCNP21] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. *ACL Findings*, pages 4275–4293, 2021.

[Tat26] Tatoeba Association. Tatoeba corpus. https://tatoeba.org, 2026. Dataset retrieved January 9, 2026. Licensed under Creative Commons Attribution 2.0 France (CC BY 2.0 FR).

[TLI+23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[Tsu11] Hideo Tsutsumi. Conversation analysis of boke-tsukkomi exchange in japanese comedy. *New Voices*, 5:147–173, 2011.

[WS19] Orion Weller and Kevin Seppi. Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*, 2019.

[YA18] Motoki Yatsu and Kenji Araki. Comparison of pun detection methods using japanese pun corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3602–3605, 2018.

[YLDH15] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2367–2376, 2015.