Understanding Irony through Explanations and Background Knowledge

Aaron Maladry, Cynthia Van Hee, Els Lefever, Véronique Hoste

Language and Translation Technology Team (LT3), Ghent University

firstname.lastname@ugent.be

Abstract

This article investigates the automatic explanation of irony in English tweets. The work covers the development and validation of a conceptual framework for annotating knowledge-informed explanations for figurative language as well as the training and evaluation of specialized generative models. Human judgements confirm that both fine-tuned open-source models (Llama 3) and proprietary models (GPT-4) can produce high-quality explanations, effectively incorporating relevant world knowledge. While metrics like BLUE and ROUGE do not seem to align with human judgement, we find that semantic similarity measures align well with human quality estimations. The resulting models and datasets for irony explanations, published as the iRONNIE collection, actively bridge the gap between theoretical understanding of irony and the technical innovations of the NLP domain. The models are be released to the public to facilitate a deeper linguistic analysis of world knowledge involved in understanding irony on social media in future work.

Keywords: Explanation generation, world knowledge, irony, sarcasm

1. Introduction

Irony and sarcasm have long been the subject of investigation within linguistics, pragmatics, and psychology, resulting in a variety of theories that attempt to capture their meaning, function, and impact. Figurative language is historically framed in relation to Grice's maxims of communication (Grice, 1975), which prescribe that communication should be truthful, informative, relevant, and clear. When these maxims are deliberately and recognizably flouted, speakers indicate that their words are not intended to be taken literally. Metaphors achieve this by describing something in terms of what it is not, thereby bending the maxim of truthfulness. Hyperboles and understatements distort degree: they exaggerate or downplay a characteristic, such as size, beyond realistic measures. Verbal irony differs from these forms in that it relies on incongruity between literal expression and reality, conveying a meaning that often contradicts the surface content (Sperber and Wilson, 1986). Other forms of irony do not invert meaning in this way, but instead highlight contrasts in different domains. Situational irony arises when real-world events play out in ways that contradict common expectations, such as when a police officer is arrested for theft. The irony lies not in words but in the tension between what is expected and what actually occurs. Dramatic irony is rooted in narrative and performance: it emerges when an audience holds knowledge that characters do not, creating a situation where the audience realizes the tragic implications of a character's actions long before they do.

While situational and dramatic irony are culturally and narratively important, most linguistic and computational research has focused on verbal irony, given its strong departure from literal meaning and its central role in everyday communication. To attain verbal irony, a negative situation can unexpectedly be described as positive, creating an evaluative incongruity (or clash) that listeners interpret as ironic (Riloff et al., 2013). This form of irony, which can be referred to as "irony by clash" (Van Hee et al., 2016), is relatively explicit, as both positive and negative sentiments are overtly present. However, irony can also be more subtle. For instance, when a positive sentiment is expressed and the negative nature of the situation must be inferred. In these cases, listeners must rely on their world knowledge and empathic perspective-taking to recognize that the speaker's intended meaning differs from the literal interpretation (Sperber and Wilson, 1986). The ability to recognize such subtleties aligns irony closely with humour (Loakman et al., 2023), but also reveals its complexity.

Beyond its linguistic characteristics, irony has been widely studied for its psychological and social functions. While it is often associated with indirect criticism or sarcasm (Brown and Levinson, 1987), irony can also serve positive communicative goals, such as reinforcing social bonds, mitigating face-threatening acts, defusing tension, expressing empathy, or even functioning as a coping mechanism to process negative emotions (Pfeifer and Pexman, 2023). Despite the longstanding interest in irony, its psychological, neurological, and linguistic mechanisms remain only partially understood. Re-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 16–28
16 May 2026.

searchers have identified typical cues, such as the use of hyperbolic statements (Kreuz and Roberts, 1995), simile (Hao and Veale, 2009), or rhetorical questions (Kreuz et al., 1999), but there is no comprehensive account of the types of knowledge listeners draw upon when interpreting irony.

A central challenge in irony research is the role of world knowledge and contextual assumptions. Theoretical linguistic work emphasizes that recognizing ironic intent often relies on access to shared or mutually assumed knowledge, including real-world facts, cultural conventions, and beliefs about the speaker's perspective (Lewis, 1969; Schiffer, 1972). Although it is widely acknowledged that understanding irony requires knowledge about real-world situations, conversational norms, and emotional expectations, it remains unclear which types of knowledge are typically invoked. This gap has also received limited attention in computational studies. Although irony has been a popular research topic, the primary focus has been on benchmarking, training systems to detect whether a text contains irony (Saroj and Pal, 2024; Basile et al., 2024).

To advance the study of irony, we explore methods that can scale up the production of meaningful irony explanations, with the final goal to enable further analysis of the world knowledge needed to understand irony. To make this possible, we engage human experts to construct a set of exemplar explanations based on world knowledge, which are subsequently used to train and evaluate generative models. In this study, we evaluate (1) whether generative models are capable of producing knowledge-informed explanations for ironic texts, (2) how these explanations compare to human-produced explanations, and (3) to what extent automatic metrics for explanation quality align with human judgments. All models and data are published as the iRONNIE collection.

2. Related Work

2.1. Explanation Generation and Reasoning

With the growing success of large generative models, explanation generation and reasoning tasks have become increasingly central to NLP research. While encoder-only models are traditionally used for classification tasks with feature-based explanations, modern methods are shifting towards generative models such as decoder-only architectures that can generate natural language explanations. This trend reflects a broader shift from post-hoc model interpretability methods towards generating explanations that are directly expressed in natural language, making them more accessible to human users. Importantly, explanation generation and model interpretability, although related, pursue distinct objectives. Model interpretability aims to make the model's internal mechanisms transparent, while explanation generation focuses on providing human-readable justifications, without necessarily aligning with the true reasoning process of the model. This distinction is crucial for generative models, whose token-by-token generation complicates efforts to study their inner workings.

To mitigate this, researchers are adapting tools from explainable AI such as SHAP (Lundberg and Lee, 2017), to the text generation setting (Enouen et al., 2024). SHAP assigns importance scores to input features (e.g., words) based on how their presence or absence affects model predictions, relying on game-theoretic principles. When applied to autoregressive generation, SHAP can identify the words that have the greatest impact on the generation of specific tokens. However, these token-level importance scores remain superficial: while they indicate *which* words matter, they do not reveal *why* they matter, especially when explanations rely on implicit knowledge or reasoning that is not mentioned explicitly in the input text.

Chain-of-Thought (CoT) prompting (Wei et al., 2022) offers an alternative strategy by asking models to reason step-by-step, decomposing the task into intermediate natural language steps before arriving at a final answer. While CoT improves performance on many classification tasks, it does not guarantee that generated explanations faithfully represent the model's true decision process (Atanasova et al., 2023; Turpin et al., 2024). This misalignment between explanation and model reasoning has motivated work on faithfulness and consistency metrics (Lanham et al., 2023; Parcalabescu and Frank, 2023). While we recognize that it is important that generated explanations reflect the models internal process,we argue that before concerns of faithfulness arise, models must first generate explanations that are meaningful, informative, and intuitive to human users, particularly when the task relies on world knowledge.

2.2. Explanation Generation for Figurative language

Within the field of figurative language processing, explanation generation is still in its early stages. For figurative language as a whole, Chakrabarty et al. (2022) introduce the FLUTE dataset, which covers metaphors, hyperboles, similes, and irony. Their approach involves constructing paraphrases to expose the intended meaning of figurative expressions and generating explanations framed as entailments or contradictions of these paraphrases via NLI models. However, FLUTE lacks human

standard explanations and applies a generic explanation format, without recognising the role of world knowledge or incorporating its explicit inclusion as a criterion for explanation quality. Subsequent work extends FLUTE to the multimodal domain (Saakyan et al., 2024) incorporating images into figurative language interpretation. Yet, even in this expanded setting, explanations remain limited to validating or contradicting paraphrases without exploring deeper pragmatic or world-knowledge-based reasoning. Focusing on multimodal sarcasm, Desai et al. (2022) produce explanations for multimodal sarcasm using a fine-tuned BART model, and Yi et al. (2025) make use of more recent decoder models in a zero-shot setting, however, both works tend to reduce paraphrases of the intended meaning and do not address the implicit assumptions or background knowledge required to understand sarcasm.

In parallel, explanation generation has also been explored in the context of abusive language detection where retrieval-augmented generation (RAG) methods have been employed to enhance the detection of hate speech, including sarcastic expressions. Di Bonaventura et al. (2025), for instance, combined knowledge from diverse databases to enrich the system input and improve hate speech classification. Notably, the authors demonstrated that instruction-tuned models such as FLAN-T5 (Chung et al., 2022) and Llama 2 (Touvron and et al., 2023) outperform proprietary models like GPT-3.5 Turbo in generating explanations when enriched with world knowledge.

While the aforementioned studies demonstrate the relevance of explanation generation for figurative language, their primary focus lies in creating datasets for benchmarking model performance. A notable limitation is that **current "explanations" are limited to paraphrasing** and do not reveal the underlying reasoning. The main contributions of this work lie in approaching explanation generation for irony as a distinct research objective, **focusing on inferring world knowledge** from large language models to deepen our understanding of figurative speech like irony. With this goal in mind, we develop and release datasets and models to set a new standard for explanation generation for figurative language.

3. Experimental Setup

3.1. Human-Written Irony Explanations

To investigate whether generative models can meaningfully explain ironic texts, we first construct a set of human-authored explanations. These serve a dual purpose in our study. First, they are used to train the models and provide high-quality, illustrative examples of how world knowledge can be used to explain irony. Second, they serve as a reference point in our evaluation: we compare system-generated explanations against human-written ones in a controlled human evaluation setup. Including this reference point in evaluation enables us to assess not only how well models perform, but also how human explanations are perceived under the same evaluation conditions.

This human reference was created by two annotators who explained the irony in a subset of 130 ironic tweets from the SemEval-2018 Task 3 corpus (Van Hee et al., 2018). Both annotators are trained linguists with a strong command of academic English. The explanations are written in full text, without imposing length limitations. To ensure the completeness of the explanations, annotators were instructed to approach the task as if they were explaining the irony to a child.This encouraged them to break down complex ideas into simple, understandable terms, ensuring that each explanation was comprehensible.
The recognition of irony often hinges on implicit, shared knowledge that is not directly stated in the text. To account for this, we introduced a second step in the annotation process: after writing the initial explanations, annotators were asked to systematically isolate and explicitly itemise the knowledge required for understanding the irony. Example 3.1 illustrates a complete annotation, according to the annotation guidelines described as a technical report (Maladry et al., 2025).

Example 3.1 `Ironic tweet`*: Loooovvveeeeeee when my phone gets wiped*
`Explanation`*: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.*
`Background knowledge`*:*

1. *When a phone gets wiped, all personal data and information is lost.*
2. *People do not like losing access to their personal data and pictures on their phone.*

Through this two-step annotation process, we collected (1) a set of detailed irony explanations and (2) systematic listings of the relevant world knowledge for each instance. The two independent perspectives for each of the 130 tweets resulted in total in 260 human-written explanations.

3.2. Methodology for Modelling Explanation Generation

For **model development**, we included a proprietary model, GPT-4-turbo (gpt-4-turbo-2024-04-09), and

two open source models: Mixtral (Jiang et al., 2024) and Llama 3 (Dubey, 2024). To establish an upper bound for the performance of generative models, we opted for larger versions of the models (Mixtral 8x7b and Llama 3 70b), as they generally outperform their smaller counterparts. Given the computational cost, restricted access, and the opacity of their internal workings, we did not fine-tune the proprietary models but included three in-context examples in the prompt. Open source models, by contrast, offer greater flexibility in terms of training and fine-tuning and can be used freely by other researchers. This allows us to employ supervised fine-tuning on the models with QLoRA (Hu et al., 2021; Dettmers et al., 2023) and a single in-context example to illustrate the expected output format and clarify the goal of the task. Fine-tuning was performed over 10 epochs, using the following parameters: 128 adapter ranks, an alpha value of 16, with an effective batch size of 4, a learning rate of 5e-5, and AdamW as the optimiser[1]. For the GPT models, we default temperature and context parameters are used to replicate the average user setting.

Early testing indicated that performing both explanation generation and knowledge extraction at once was too demanding for the systems and resulted in too noisy outputs. To counter this issue, we split the annotation into two subtasks: (1) explanation generation (prompts in Appendix A and (2) knowledge extraction (prompts in Appendix B), with equal methodology and parameters. For knowledge extraction (2), the model input includes the ironic tweet with explanation, either human-written or generated in task (1).

For training, the dataset of 130 explained tweets was split into a test set of 80 samples and a train set of 50 samples. We specifically opted for a larger test set – to improve the robustness of our evaluation – and a smaller train set –because we assume that large generative already encoded the relevant information and only need to be fine-tuned to understand the scope of the task and the expected output format. The train set of 50 samples and test set of 80 samples are available at Amala3/iRONNIE_train and Amala3/iRONNIE_EN_evalset.

3.3. Qualitative Evaluation for Explanation Generation

In our evaluation setup, we conducted a manual evaluation of the explanations provided by the two fine-tuned open source generative models (Llama 3 and Mixtral), the zero-shot GPT-4 Turbo model with in-context examples and the two human explainers. For clarity, we will refer to the individuals who wrote the human explanations as "explainers" and we will use the same term to describe the generative models tasked with generating explanations. We do not automatically assume that human explanations are better than system-generated, and, therefore, refer to the human-written explanations as `human` standard instead of `gold` standard. This allows for consistent terminology across both human and system-generated outputs. For this qualitative assessment, each of the 80 tweets was evaluated by five humans and all five explanations for the tweet were shown simultaneously. For consistency, all five explanations for each tweet were shuffled, anonymised, and underwent minimal manual cleaning before being presented to the annotators.

The **qualitative evaluation framework** consists of three components, based on related work (Desai et al., 2022; Saakyan et al., 2025): (1) indicating for each explanation whether the explanation is **adequate**, meaning it reasonably explains why the text is ironic, (2) indicating whether the explanation is written by a **human or a system** and (3) **ranking** the explanations from **best to worst** based on their personal preferences. Detailed annotation guidelines for explanation evaluation as well as examples are relayed in a technical report (Maladry et al., 2025).

	Krippendorff's α	Weighting
Adequacy	0.7092	Binary
Human-likeness	0.6662	Binary
Rank Scoring	0.7200	Linear

Table 1: Observed agreement between raters for our three evaluation criteria, along with the weighting schemes used to account for disagreement.

For each of the three evaluation criteria, we first determined the reliability of the evaluating scheme by calculating inter-annotator agreement using Krippendorff's α (Krippendorff, 2011).[2] The first two annotation levels, *adequacy* and *human-likeness* are binary annotations for each explanation. For the third annotation level, we consider the *rank* (first, second, third, etc.) as a scoring mechanism, with the highest-ranked explanation receiving 4 points, the second 3, and so on and the lowest-ranked explanation receiving 0 points. If an explanation is ranked first (4 points) by one annotator and second to last (1 point) by another, the label distance for the explanation is 3 points. Relative to the maximum label distance of 4, this means the error for this specific example has a weight of 0.75 (3/4). With five

[1] Example scripts for fine-tuning and inference are available on Github at aMala3/IronyExplanations

[2] All IAA scores are calculated with the irrCAC (https://github.com/kgwet/irrCAC) package for statistical agreement measures in R.

	Adequacy			Human-likeness			Ranking Score		
	Avg.	Δ	Sig.	Avg.	Δ	Sig.	Avg.	Δ	Sig.
GPT-4 Turbo	0.72	+0.11	**	0.05	-0.42	***	2.96	+1.07	***
human2	0.66	+0.06	–	0.47	-0.01	–	2.49	+0.60	***
→ human1	0.60	0.00	–	0.48	0.00	–	1.89	0.00	–
Llama 3	0.53	-0.08	*	0.48	0.00	–	1.53	-0.36	***
Mixtral	0.39	-0.22	***	0.39	-0.08	*	1.14	-0.75	***

Table 2: Evaluation of Adequacy, Human-likeness, and Ranking Score for each explainer. For each criterion, the table displays the average score, the mean difference (Δ) with regards to human1, and significance w.r.t. the human standard. The asterisks mark the degrees of significance, with * indicating p<0.05, ** indicating p<0.01, *** indicating p<0.001 and a dash (–) marks no statistically significance.

evaluations for each of the 400 explanations, the resulting set comprises 2,000 annotations for which we calculated inter-rater agreement. The α-scores (Table 1) indicate that both adequacy and ranking score meet the recommended values (>=0.667) for reliability (Krippendorff, 2011), while scores for human-likeness annotations are barely short of this value. Overall, these scores support the **acceptable reliability of the annotation scheme**.

Table 2 reports **results** across three dimensions: adequacy, human-likeness, and ranking for both system-generated and human-written explanations.

For **Adequacy**, no statistically significant differences are found between the two human explainers. In contrast, significant differences emerge when comparing the systems to the human explainers. Whereas GPT-4 scores higher on adequacy than both humans, Llama scores slightly lower. Mixtral explanations exhibit the greatest divergence from human explanations, scoring 22% lower on adequacy compared to human1, a difference that is highly statistically significant.

For **Human-likeness**, we observe no significant difference in scores between the human explainers, importantly, nor between the humans and our fine-tuned Llama 3 model. This suggests that human annotators were unable to distinguish between human-written explanations and those generated by the fine-tuned Llama model. Surprisingly, despite its low adequacy scores, Mixtral is often perceived as human-like, with a difference of only 7% (Δ) and limited statistical significance (p<0.05). In contrast, GPT-4's explanations are readily identified as non-human. A closer look at the explanation lengths further supports these observations. The average word count for explanations from the fine-tuned Llama 3 (41 words) and Mixtral (43 words) models closely aligns with that for the human explanations, with 45 and 43 words on average. In stark contrast, GPT-4 Turbo's explanations -averaging 105 words- were exceedingly longer than typical human responses and were almost always recognized as non-human.

For the **Rank** scores, we find that all score differences are highly statistically significant and that the explanations of GPT-4 were often preferred over the explanations written by humans. Among the human annotators, the explanations written by human2 seem to be preferred over those by human1. The score differences are smallest between human1 and fine-tuned Llama, demonstrating the similarity of these two explainers.

Although some scores may appear low in absolute terms, with for instance 57.9% of the explanations judged adequate and 37.3% labelled human-like, this reflects the comparative nature of the task. Presenting five explanations simultaneously encourages stricter evaluation, as annotators implicitly contrast responses against one another. This setup may lower raw ratings, especially for human-likeness, where even minor stylistic deviations from human norms stand out. For this reason, we interpret system performance relative to a human expert baseline, rather than in isolation.

In summary, GPT-4 explanations were most preferred and were generally rated as most adequate, while fine-tuned Llama 3 matched human-level performance on human-likeness. Mixtral lagged behind across all dimensions.

3.4. Qualitative Evaluation of Knowledge Extraction

Before evaluating the content of the knowledge extraction task, we inspect the length of the extracted knowledge. For now, we treat all extracted knowledge as a single text, avoiding the need to align individual knowledge items or account for variations in the number of produced knowledge items.

Whereas the Mixtral output for explanation generation could be cleaned manually with relative ease, the output for knowledge extraction was too noisy, inconsistent and repetitive to be used reli-

ably on a large scale. Therefore, we continued the knowledge experiments with Llama 3 and GPT-4[3]. Table 3 presents the average word counts of background knowledge items extracted from explanations authored by each source (rows), using different extraction systems (columns). For example, 31 indicates the average word count of knowledge items extracted by GPT-4o from explanations written by `human1`, 19 is the average word count of knowledge items human annotators extracted from the explanations written by `human2`. Note that there is only one human column because humans only extracted the background knowledge in their own explanation.

	GPT-4o	Llama 3	Human
`human1`	31	27	22
`human2`	30	24	19

Table 3: Average word lengths of background knowledge extracted from explanations by different sources (rows), using different extractors (columns).

We observe that GPT-4o consistently produces longer knowledge items, regardless of the explanation source, averaging 31 words for `human1` and 30 for `human2`. The two human explainers, in contrast, extract more concise knowledge (19–22 words), with Llama 3 falling somewhere in between. This mirrors trends in explanation verbosity, where GPT-4o also generated significantly longer responses.

Compared to explanation generation, knowledge extraction may seem less interpretive and more objective, but its evaluation remains challenging. For instance, a single piece of background knowledge can be rephrased as a self-contained statement or split into multiple shorter ones. Similarly, related concepts may be consolidated into a broader knowledge unit. These variations, while often semantically equivalent, can significantly impact automated precision and recall scores.

To assess extraction performance more directly, we conduct a manual validation on a subset of 50 explanations provided by `human1`, evaluating knowledge extracted by Llama 3 and GPT-4o. For this validation, we identify **whether all essential background knowledge was successfully extracted** from the explanation (in the sense of recall), and **whether all of the produced knowledge was present in the explanation** (in the sense of precision). Importantly, we do not compare extracted knowledge to a predefined human standard. Instead, each knowledge item is evaluated independently against its source explanation through manual inspection. As a result, these metrics reflect a looser, more interpretive approximation of traditional precision and recall, which we refer to as `soft recall` and `soft precision`. Although some knowledge items are present in the explanations, and are considered correct for soft precision, they still deviate slightly from ideal precision. Therefore, we also analyse two common deviations from ideal precision: **content drift**, where the extracted item differs slightly in meaning (e.g., due to over-generalization or added specificity), and **reasoning**, where the item goes beyond background knowledge to include evaluative or inferential content that should belong in the explanation rather than the extracted knowledge. Example 3.2 shows content drift, where the extracted knowledge is not explicitly present in the explanation, but is implied in the part in bold. Example 3.3 shows a reasoning deviation: the extracted knowledge contains reasoning (in bold), making the contrast explicit between the background knowledge. Although this reasoning should be included in the explanations, it should not be extracted as background knowledge.

Example 3.2

`Tweet:` *@user I don't recall dalai lama talking about busty girls and celebs but apart from that*

`Explanation:` ***The Dalai Lama is a religious leader, which are supposed to be moral guides**. Talking about "busty girls" and celebrities is considered rather shallow and would be inappropriate for a religious leader. Therefore, it is highly unlikely that this ever happened.*

`Extracted Knowledge:`

(1) The Dalai Lama is expected to maintain a certain level of moral and spiritual integrity.

Example 3.3

`Tweet:` *Throwing up on Christmas morning is my ideal way of spending it.*

`Explanation:` *Throwing up is a very unpleasant experience that is the result of being sick. When people are sick, they likely stay in bed. On Christmas day or morning, people like celebrating instead, which you cannot do when you are sick.*

`Extracted Knowledge:`

*(1) Christmas is a time for celebration and spending time with others, **rather than being bedridden**.*

As shown in Table 4, both systems achieve strong performance. GPT-4o extracts all relevant items (100% recall) with high precision (95.5%), and only 6.3% of its extractions exhibit minor content drifts. Llama 3 achieves slightly lower recall (95.9%) and precision (90.3%), with a higher proportion of meaning shifts (16.1%). While both models occasionally include reasoning content, these cases are generally well grounded in the explanation and do not represent major departures from the task definition.

[3]As we are now generating text for 400 samples (extracting knowledge from 5 explanations for each of the 80 texts), we opted for GPT-4o instead of GPT-4 Turbo to save computational resources.

	Llama 3	GPT-4o
Recall	95.88%	100.00%
Precision	90.29%	95.49%
Content Drift	16.13%	6.30%
Includes Reasoning	10.75%	12.60%
Total Items	103	133

Table 4: Soft recall and soft precision of extracted knowledge items, based on manual evaluation of 50 explanations from human1. Rather than comparing against a predefined human standard list, each extracted item is assessed relative to its source explanation. Content drift and reasoning refer to knowledge items that are considered correct for soft precision, but exhibit deviation from the ideal desired output.

Overall, both Llama 3 and GPT-4o accurately extract relevant background knowledge. While GPT's outputs are more verbose and sometimes fragmented into smaller components, this verbosity does not indicate hallucination or content drift. Whereas GPT produces slightly more reasoning items, Llama's extractions remain closer in style and length to human-produced items, but they can exhibit slight content drift.

3.5. Automatic Similarity Measures for Evaluation

While our primary evaluation relies on human judgments, this approach requires two stages of manual annotation: first to produce human-written explanations, and second to evaluate generated (and human) explanations. This process is resource-intensive and limits the number of systems that can be compared, particularly due to the cognitive load on evaluators when reviewing multiple explanations simultaneously. In this section, we explore whether automatic similarity metrics can serve as a viable alternative for future evaluation. If such metrics reliably reflect differences in explanation quality, especially in alignment with human judgments, they could enable more scalable and efficient comparison of additional systems or configurations. To test this, we investigate whether established lexical and semantic similarity metrics can distinguish between stronger and weaker explainers or knowledge extractors.

As **lexical metrics**, we employ traditional BLEU (Papineni et al., 2002) and ROUGE scores (Lin, 2004), which are often used to evaluate the quality of machine translation and summarization. Additionally, we calculate vocabulary similarity using the Jaccard index (intersection over union) after lemmatizing all words in the explanations. For the semantic metrics, we leverage BERTscore (Zhang et al., 2020), a widely accepted metric for machine translation evaluation that measures semantic similarity at the token level based on contextual embeddings, and SemScore (Aynetdinov and Akbik, 2024), a novel metric for assessing the broader contextual meaning using sentence-level embeddings from transformer models (Reimers and Gurevych, 2019).

As outlined in our annotation guidelines, human explainers wrote explanations that included both the necessary background knowledge and the reasoning steps they deemed necessary to understand the irony in a text. In the second annotation step, they separated the knowledge from the reasoning process. For this similarity analysis, we first examine the similarities between the full explanations and then focus on the isolated background knowledge.

When **comparing human and generated explanations** (Figure 1), none of the three metrics for lexical similarity (BLEU, ROUGE and Jaccard vocabulary similarity) exceed 20%, indicating that the word choices vary substantially between explanations. This variation is observed not only in the machine-generated explanations, but also holds true for the "inter-human" comparison of the two human explanations. Despite these lexical differences, the BERTscores indicate a higher degree of token-level semantic similarity, suggesting that, even though different words are used, they are semantically aligned. Similarly, SemScores confirm that sentence-level meaning also remains consistent across human and machine explanations.

For the automatic **evaluation of knowledge extraction**, we focus on the two best-performing models, Llama and GPT-4o, as knowledge extracted by Mixtral required too much manual cleaning for this task. With these models, we automatically extracted knowledge from the explanations of human 1 and human 2, and calculated the similarity compared to the human-identified knowledge. Since both human explainers only extracted knowledge from their own explanations, we are unable to present inter-human similarities for this task. However, this comparison enables us to investigate how well the models extract knowledge from the same explanation.

	BLEU	ROUGE	Jacc.	BERT	SEM
GPT	0.27	0.36	0.33	0.81	0.67
Llama	**0.37**	**0.57**	**0.47**	**0.87**	**0.85**

Table 5: Similarity scores between human-written knowledge and extracted knowledge using GPT-4o and Llama 3, compared to the human1 standard.

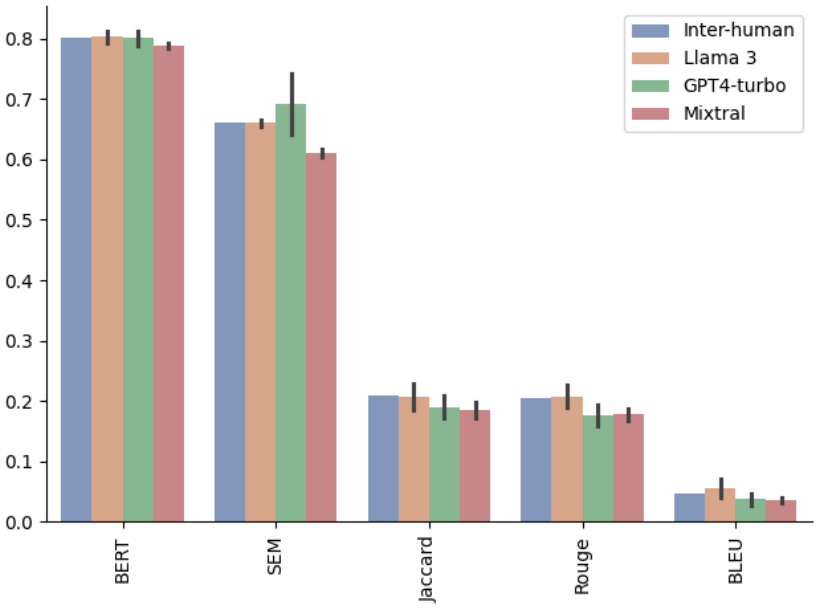

Figure 1: Full explanation similarities for two humans and all systems. Similarities were calculated for both annotators individually and then averaged.

As shown in Table 5, Llama 3 outperforms GPT-4 on all similarity metrics, both lexical and semantic measures. Compared to the task of explanation generation, the similarities observed through the lexical measures (BLEU, ROUGE, Jaccard vocabulary similarity) are much higher for knowledge extraction, suggesting that the phrasing is generally quite similar. Although manual evaluation of GPT-4o indicated that the model could extract a lot of relevant knowledge, the semantic similarity scores are quite low for the model. We hypothesize that this could be connected to the fact that GPT-4 included more reasoning in its knowledge items. Although Llama produced more content drift, the semantic similarities do not seem to capture this difference in nuance.

4. Conclusion

In this study, we investigate how well automatic systems can explain irony in English tweets and whether or not language models can produce explanations that contain relevant world knowledge that is not explicitly mentioned in the input text. In doing so, this investigation evaluates explanation quality, i.e. whether generative language model can provide meaningful explanations that are intuitive to humans and align with linguistic theory about irony, as opposed to evaluating the faithfulness of explanations, which we leave for future work. The models and datasets produced for this work are available on Hugging Face as the iRONNIE collection. Based on our analysis, we present the following insights:

(1) Our annotation scheme, based on adequacy, human-likeness, and full-scale ranking, provides a **consistent and reliable framework for evaluating explanation quality** in the context of irony detection.

(2) **GPT-4 generates explanations that are preferred over human explanations**, even though the extensive explanation length makes the model easily discernable from human explanations.

(3) With appropriate fine-tuning,**Llama 3 produces explanations that are often indistinguishable from human-written ones**, although they are rated slightly lower in adequacy and ranking.**Semantic similarity** metrics, particularly SemScore, **seem to align with human judgments**, offering a promising avenue for automatic evaluation of explanation quality.

(4) Fine-tuned open source and proprietary models are capable of reliably extracting world knowledge for explaining irony, demonstrating that generative **AI can make relevant world knowledge explicit**.

5. Acknowledgements

This work was supported by Ghent University under grant BOF.24Y.2021.0019.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Limitations

This study has several limitations that should be acknowledged. First, the human evaluation relies on a limited number of annotators and was conducted on modest test sets. While we adopted rigorous annotation procedures and achieved strong inter-annotator agreement, a broader evaluation effort would be needed to confirm the generalizability of our findings across diverse datasets, user groups and languages. It is widely known that generative models perform significantly better for English than they do on other languages, so even for the same task, future work should explore if model capabilities transfer well across languages. This is particularly relevant when analysing the role of world knowledge, which is captured implicitly and may required a vast amount of data to be learned.

While this study finds that automatic evaluation metrics align with human judgement, this finding should be further investigated before it can be confirmed. Similarly as the results for explanation generation, whether or not these metrics can also be used to evaluate languages other than English remains an important research question.

Our aim in this work was not to exhaustively optimize every model configuration, but rather to evaluate the capabilities of current models, particularly large language models, for the task of irony explanation and knowledge extraction. As such, we focused on representative configurations rather than fine-tuning all hyper-parameters. More extensive optimization, especially for smaller open-source models, is complicated by the fact that automatic evaluation metrics do not consistently align with human judgments, making it difficult to systematically assess model improvements without costly manual validation.

6. Bibliographical References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. *ArXiv*, abs/2305.18029.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity.

Valerio Basile, Silvia Casola, Simona Frenda, and Soda Marem Lo. 2024. Perseid-perspectivist irony detection: A calamita challenge.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2025. From detection to explanation: Effective learning strategies for LLMs in online abusive language research. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2067–2084, Abu Dhabi, UAE. Association for Computational Linguistics.

et al. Dubey, Abhimanyu. 2024. The llama 3 herd of models.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. 2024. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13984–14011, Bangkok, Thailand. Association for Computational Linguistics.

H. P. Grice. 1975. *Logic and Conversation*, pages 41–58. Brill, Leiden, The Netherlands.

Yanfen Hao and Tony Veale. 2009. Support structures for linguistic creativity: A computational

analysis of creative irony in similes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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

Roger J Kreuz, Max A Kassler, Lori Coppenrath, and Bonnie McLain Allen. 1999. Tag questions and common ground effects in the perception of verbal irony. *Journal of Pragmatics*, 31(12):1685–1700.

Roger J Kreuz and Richard M Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and symbol*, 10(1):21–31.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1:25–2011.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Tom Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *ArXiv*, abs/2307.13702.

David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Aaron Maladry, Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2025. Guidelines forWriting and Evaluating Explanations for Irony in Tweets.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Letitia Parcalabescu and Anette Frank. 2023. On measuring faithfulness or self-consistency of natural language explanations. In *Annual Meeting of the Association for Computational Linguistics*.

Valeria Pfeifer and Penny Pexman. 2023. When it pays to be insincere: On the benefits of verbal irony. *Current Directions in Psychological Science*, 33.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Understanding figurative meaning through explainable visual entailment.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment.

Anita Saroj and Sukomal Pal. 2024. Ensemble-based domain adaptation on social media posts for irony detection. *Multimedia Tools and Applications*, 83(8):23249–23268.

Stephen R Schiffer. 1972. Meaning.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Hugo Touvron and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1794–1799.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Peiling Yi, Yuhan Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A. Cleaning Examples

Below, we exemplify the manual cleaning of the generated output of our fine-tuned Mixtral model. We present the output before cleaning in Example A.1, including input, instruction tuning tokens and a hallucinated second tweet with explanation. The same output after cleaning, as presented to the annotators, is shown in Example A.2.

Example A.1 `Mixtral output before cleaning:`
Text: I actually just started crying while brushing my teeth due to lack of sleep, stress and test anxiety. Today's gonna be a good one...
Explanation: [/INST] ### Explanation: Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.
[Hallucinated] Text: 25 chance of rain, I love it when my wedding gets ruined!
Explanation: [/INST] ### Explanation: [/INST] People want their wedding day to be perfect. Rain can cause damage to the wedding dress and the hair, it can make the wedding less enjoyable and even make it impossible to have the wedding outside. As people would not like it when their wedding gets ruined by rain, the positive evaluation is not meant to be serious.

Example A.2 `Explanation after manual cleaning:`
Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.

B. Prompt Template for Explanation Generation

GPT-4 Turbo

Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.
Here are some examples of the expected output:

Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.

Example 2 -
Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous
Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.

Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.
Explanation: The stereotypical phishing email from a supposed "nigerian prince" is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).

New sample:
Text: {text placeholder}
Explanation: {system fills in explanation}

Llama 3 & Mixtral

[Human Says][BoS]
You are an expert trained in identifying irony and sarcasm in social media text and explaining the underlying reasoning. Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.
Explain why this tweet is ironic:
Text: nit pick at every little thing guys. keep doing that. that'll show the team you're behind them.[EoS]

[System Says:][BoS]
Explanation: nitpicking (pointing out a lot of minor faults) is often perceived as providing too much negative feedback. this does not help to support a team, i.e. showing the team you're behind them.[EoS]

[Human Says:][BoS]
Explain why this tweet is ironic:
Text: {text placeholder} [EoS]

[System Says][BoS]
Explanation: {system fills in explanation}

C. Prompt Templates for Knowledge Extraction

GPT-4o

You will be provided with an ironic tweet along with an explanations for why they are ironic. These explanations are based on extra textual background knowledge. Your task is to extract which knowledge is used in the explanations but which is not present in the text itself. Split up this information into multiple items where this is appropriate. Make sure that the knowledge you provide is mentioned explicitly in the explanation. Here are some examples of the expected output:

Example 1 -
Text: Loooovvveeeeeee when my phone gets wiped
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.
Background knowledge:
1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone.

Example 2 -
Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous
Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is know for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.
Background knowledge:
1. When people say that someone has "the guts to do something", it means they are fearless, confident and brave.
2. The cartoon character Scooby Doo is known for his cowardice.

Example 3 -
Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.
Explanation: The stereotypical phishing email from a supposed "nigerian prince"is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).
Background knowledge:
1. The email from a Nigerian prince or heir is a well-known stereotype of phishing.

New sample:
Text: {text placeholder}
Explanation: {explanation placeholder}
Background knowledge: {system fills in knowledge items}

Open-Source Models

[Human Says][BoS]
For each of the following input texts, we have an explanation for why they contain irony. Your task is to identify and extract any background knowledge used for the explanation.
Extract the background knowledge from this text and explanation pair:
Text: Loooovvveeeeeee when my phone gets

wiped.
Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic. [EoS]

[System Says][BoS]
Background knowledge:
1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone. [EoS]

[Human Says][BoS]
Extract the background knowledge from this text and explanation pair:
Text: {text placeholder}
Explanation: {explanation placeholder} [EoS]

[System Says][BoS]
Background knowledge: {system fills in knowledge items}