Draw Me Like My Triples: Leveraging Generative AI for Wikidata Image Completion

Raia Abu Ahmad, Martin Critelli, Şefika Efeoğlu, Eleonora Mancini, Celian Ringwald, Xinyue Zhang, Albert Meroño Peñuela

Draw Me Like My Triples: Leveraging Generative AI for Wikidata Image Completion

Raia Abu Ahmad[1], Martin Critelli[2], Şefika Efeoğlu[3,4], Eleonora Mancini[5], Célian Ringwald[6], Xinyue Zhang[7] and Albert Meroño-Peñuela[8]

[1]*Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH), Berlin, Germany*
[2]*University Ca'Foscari of Venice, Italy*
[3]*Freie Universität Berlin, Germany*
[4]*Technische Universität Berlin, Germany*
[5]*University of Bologna, Italy*
[6]*Université Côte d'Azur, Inria, CNRS, I3S, France*
[7]*University of Oxford, Oxford, United Kingdom*
[8]*King's College London, London, United Kingdom*

Abstract

Humans are critical for the creation and maintenance of high-quality Knowledge Graphs (KGs). However, creating and maintaining large KGs only with humans does not scale, especially for contributions based on multimedia (e.g. images) that are hard to find and reuse on the Web and expensive to generate by humans from scratch. Therefore, we leverage generative AI for the task of creating images for Wikidata items that do not have them. Our approach uses knowledge contained in Wikidata triples of items describing fictional characters and uses the fine-tuned T5 model based on the WDV dataset to generate natural text descriptions of items about fictional characters with missing images. We use those natural text descriptions as prompts for a transformer-based text-to-image model, Stable Diffusion v2.1, to generate plausible candidate images for Wikidata image completion. We design and implement quantitative and qualitative approaches to evaluate the plausibility of our methods, which include conducting a survey to assess the quality of the generated images.

Keywords

Generative AI, Image Generation, Automated Prompt Generation

1. Introduction

Large knowledge bases (KBs) such as Wikidata are maintained by human editors in a collaborative manner in order to provide structured data of high quality [1]. However, given the size of this platform, there is an evident problem of incompleteness that creates several content gaps [2]. We note that this is especially true for contributions based on multimedia (such as images,

Wikidata'23: Wikidata workshop at ISWC 2023
raia.abu_ahmad@dfki.de (R. Abu Ahmad); martin.critelli@unive.it (M. Critelli); sefika.efeoglu@fu-berlin.de,tu-berlin.de (Ş. Efeoğlu); e.mancini@unibo.it (E. Mancini); celian.ringwald@inria.fr (C. Ringwald); xinyue.zhang@cs.ox.ac.uk (X. Zhang); albert.merono@kcl.ac.uk (A. Meroño-Peñuela)
0009-0004-8720-0116 (R. Abu Ahmad); 0000-0002-8177-730X (M. Critelli); 0000-0002-9232-4840 (Ş. Efeoğlu); 0000-0001-9205-3289 (E. Mancini); 0000-0002-9232-4840 (C. Ringwald); 0000-0002-9232-4840 (X. Zhang); 0000-0003-4646-5842 (A. Meroño-Peñuela)

audio, and video) since it is difficult for editors to find such high-quality contributions on the Web, and even more difficult and expensive to create them from scratch.

In this work, we examine the problem of missing images for a specific class of Wikidata entities: **fictional characters**. We motivate this choice by the fact that querying Wikidata shows that only 7% out of the 83.7K instances of the fictional character class, including its sub-classes, have an image [1]. It is important to note that this class was specifically chosen due to ethical and privacy concerns, as other classes (e.g. person) can have detrimental consequences if automatic images are created to portray them. Although alternative methods for finding images for this class of entities exist (e.g. fan-created images), they are unreliable in terms of the objective representation of characters and demand thorough and manual research by editors to make sure that they correctly align with Wikidata's knowledge about each entity.

We propose a novel method of leveraging knowledge from Wikidata triples about each fictional character entity in order to create a representative image for it using generative artificial intelligence (AI) models. This is done by (1) extracting triples from Wikidata entities, (2) creating English prompts to be fed into a generative text-to-image model such as Stable Diffusion [3], and (3) generating a representative image of the character that could potentially be used on Wikidata. We investigate the effectiveness of this approach by generating four different types of prompts in English, including using triple verbalisation with large language models (LLMs), for each character and comparing the resulting images. We evaluate our approach based on a ground-truth dataset that consists of fictional characters which already have an image on Wikidata. We select different metrics of automatic image comparison to measure how similar each generated image is to the ground-truth one. Additionally, since automatic measures for image comparison are limited, we conduct a human evaluation survey in which we ask participants to evaluate image similarity.

Our work addresses the following research questions (RQs):

- **RQ1**: To what extent can different types of prompts based on triples be used in text-to-image models to produce high-quality images?
- **RQ2**: To what extent can the output of generative AI be used for Wikidata image completion?
- **RQ3**: How can generative text-to-image models be evaluated?

To the best of our knowledge, no previous study has explored the realm of using Wikidata as a source for creating prompts for generative text-to-image models. Our work [2] offers the following contributions:

- A framework that generates prompts for a text-to-image model (Stable Diffusion v2.1 [3]) with different levels of structure and natural language text based on Wikidata triples.
- A dataset of generated images for fictional characters extracted from Wikidata that can potentially be used by editors for image completion.
- An evaluation strategy showing evidence of relevancy and adequacy of using AI-generated images for our use case.

[1] This query was performed in June 2023.

[2] The project and dataset are available at https://github.com/helemanc/gryffindor and at https://huggingface.co/gryffindor-ISWS, respectively.

[3] The model card is at https://huggingface.co/stabilityai/stable-diffusion-2-1

2. Related Work

Generative AI: Current groundbreaking advances in AI enable machines to generate novel and original content based on textual prompts. Such generative applications include text-to-text [4], text-to-image [5, 6], and even text-to-music [7]. Generally, these models can capture complex patterns from the input text and produce coherent outputs. A recent survey [8] shows that text-to-image applications specifically have been emerging since 2015, when AlignDRAW [9] pioneered the field by leveraging recurrent neural networks (RNNs) to encode textual captions and produce corresponding images. Since then, end-to-end models started leveraging architectures such as deep convolutional generative adversarial networks (GANs) [10, 11, 12], autoregressive methods [13, 14, 15], latent space models [16, 17, 18], and the current state-of-the-art diffusion-based methods [19, 20, 21].

Prompt Engineering: Because of the aforementioned advances, a novel area of *prompt engineering* has emerged, in which humans interact with AI in an iterative process to produce the best prompt (i.e. textual input) for a specific desired output [22]. Recent work has shed light on prompt engineering for AI art specifically [23], concluding that simple and intuitive prompts written by humans are not enough to get desired results. Rather, writing good prompts is a learned skill that is enhanced by the usage of specific prompt templates[4] and modifiers [24].

Automatic Prompt Generation: When it comes to automatic prompt generation, previous studies tend to investigate using LLMs to construct prompts using techniques such as text mining, text paraphrasing, and data augmentation [25]. However, to the best of our knowledge, no work has touched upon using large KBs such as Wikidata for prompt engineering and generation.

3. Proposed Approach

We conduct our study on instances of the class designated as *fictional character with Q95074 item ID*. The initial stage of our approach involves the extraction of relevant triples pertaining to a specific entity. Subsequently, these triples are used to form various types of prompts in English, functioning as inputs to Stable Diffusion [21], a text-to-image AI model. We generate different types of prompts related to the triples, including a *verbalised triples prompt* which uses the T5 language model fine-tuned on the WDV dataset [26]. This verbalisation model converts triples into fluent language [27]. The ultimate goal is to generate suitable images that can serve as accurate visual identifiers for their corresponding Wikidata entities. This pipeline is shown in Fig. 1.

3.1. Triple Extraction

Generating an image for a specific character requires a description that can be gathered from its related triples. In Wikidata, these can be obtained through SPARQL queries [5], yielding all triples with the character as the subject. Moreover, since properties and entities might be represented

[4] https://sweet-hall-e72.notion.site/A-Traveler-s-Guide-to-the-Latent-Space-85efba7e5e6a40e5bd3cae980f30235f

[5] All the SPARQL queries with detailed explanations are available at https://github.com/helemanc/gryffindor/blob/main/src/data-collection/wiki_query_service.py.

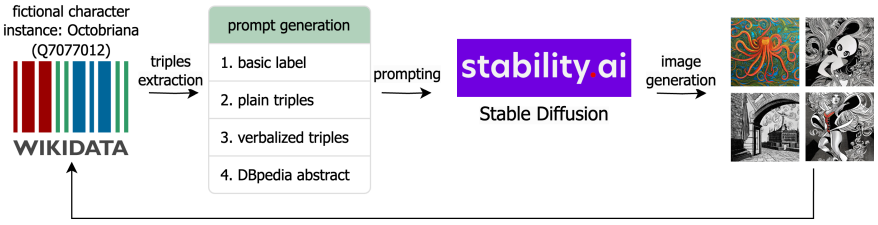

Figure 1: The pipeline of our proposed KG completion process.

by item IDs and property IDs, we also extract and translate each triple from these IDs to their corresponding labels when the triple has an entity and property ID.

3.2. Prompt Generation

In order to investigate if prompts closer to natural language work better at generating better images, we generate four distinct prompts for each character. The first three are created utilising the set of triples that have been extracted for the respective entity from Wikidata, while the last prompt utilises English DBpedia abstracts. These prompts are defined as follows (see Table 15 for prompt samples of an entity):

1. *Basic Label*: This prompt merely employs the "label" that Wikidata assigns to its entities.
2. *Plain Triples*: This prompt is derived by concatenating the subject, predicate, and object of a triple to form a single sentence, utilising all available triples linked to a specific entity. Notably, sentences generated from plain triples may lack proper structure and grammar.
3. *Verbalised Triples*: Triple verbalisation is defined as the transformation of structured data (i.e., triples) into human-readable formats (i.e., text). These serve as a summarised paragraph of all input triples.
4. *DBpedia Abstracts*: We use DBpedia abstracts as prompts obtained by querying the English chapter of DBpedia [28]. Originally written by human editors on Wikipedia, these abstracts are automatically extracted by DBpedia, preprocessed, and shortened. Unlike previous prompt types, this is the only one originally written by a human in natural language.

When examining the triples for a single entity, we observe that triples sharing the same predicate tend to contain redundant information. As a result, prompts generated directly from these plain or verbalised triples will repetitively state the same facts. However, the "instance of" predicate seems to provide distinct information for each triple. To avoid duplicating facts in prompt types (2) and (3), we remove duplicate predicates, except for "instance of", for the input triples. Among the remaining triples that share the same predicate, we keep only the one with the longest object, since longer objects likely contain more detailed information than shorter objects with the same predicate.

3.3. Image Generation

To ensure reproducibility in image generation, we utilise Stable Diffusion version 2.1 [6], an open-source text-to-image model developed by Stability AI limited to the English language. We chose version 2.1 because it supports all input shapes up to 1024x1024 and has a better performance according to benchmark evaluation results [29].
It is important to note that this particular model has inherent limitations when it comes to generating images related to the human body. To address this issue and enhance its image generation capabilities, we employ the implementation of negative prompts that have been suggested and shared on a public GitHub repository [7]. By incorporating these negative prompts, we aim to mitigate malformations in images (e.g. crossed eyes, more than five fingers, etc.).
Moreover, since Stable Diffusion has a limitation on the number of tokens allowed in the prompt sentence(s), we embed the prompt by utilising the encoder and tokenizer from Stable Diffusion, courtesy of the Compel library [8]. The model runs positive prompts of 1500 fictional characters without existing images on Wikidata and 1500 with images on Wikidata, the latter to be used for building a ground-truth dataset for evaluation.

4. Collected Dataset

We construct an extensive dataset [9] comprising 1500 fictional characters with images, as well as 1500 fictional characters without images, which are randomly chosen from the entire set of fictional characters on Wikidata. Our motivation for collecting data on fictional characters rather than real people lies in our commitment to upholding ethical standards and safeguarding privacy. Also, there is no available dataset about fictional characters, and our data collection source codes [10] can be easily applied to different domains by changing parameter settings.
In addition, we extend our data by fetching the Wikipedia abstracts of each fictional character from the English chapter of DBpedia. Although a majority of these fictional characters lack information in DBpedia because it is constructed using English Wikipedia, this is not a problem in our case since the Stable Diffusion model can only use English text as input. Since most of the fictional characters on Wikidata (ca. 78% [11]) do not have any English Wikipedia page, we only managed to gather DBpedia abstracts for 925 fictional characters with images on Wikidata and 341 fictional characters without images (see Table 1).
By analysing basic statistics from Table 1, we directly notice a big descriptive gap in terms of triples, the number of extracted unique relations, and the length of the prompts between the two datasets we constructed. Moreover, we notice that the length of the prompt is usually the shortest for verbalised triples and the longest for plain triples. After gathering the data about the fictional characters from Wikidata and DBpedia, four different prompts are automatically

[6] Stable Diffusion v2.1 model card: https://huggingface.co/stabilityai/stable-diffusion-2-1
[7] The negative prompts: https://github.com/mikhail-bot/stable-diffusion-negative-prompts
[8] Compel encodes and decodes the portion of the prompt and available at https://github.com/damian0815/compel
[9] The entire dataset is available at https://huggingface.co/gryffindor-ISWS
[10] The data collection codes are available at https://github.com/helemanc/gryffindor/blob/main/src/data-collection/wiki_query_service.py
[11] The percentage is computed on Sept. 6, 2023

constructed by using the approach described in section 3. One example is shown in Figure 2, which depicts the ground-truth image for the character **Harlequin** with its four generated images. Based on this example, it is instantly clear that some of the prompts can produce images more similar to the ground truth.

Table 1
Statistical information about the datasets used in the evaluation of the approach.

	Characters with Images	**Characters without Images**
# entities	1500	1500
# of DBpedia abstracts	925	341
# of Wikidata triples	35 281	23 157
Average # of relation by entity	19	15
Average # of unique relation by entity	9	6
Mean tokens length of DBpedia abstracts	213	175
Mean tokens length of plain triples	321	199
Mean tokens length of verbalised triples	89	68

Figure 2: Images for the character of *Harlequin*. (a) Ground truth from Wikidata. (b) Generation from the basic label prompt. (c) Generation from the plain triples prompt. (d) Generation from the verbalised triples prompt. (e) Generation from the DBpedia abstract prompt.

5. Evaluation

In order to understand whether the generated images can plausibly be used for representing fictional characters based on their Wikidata triples, we employ two evaluation strategies. The first is an automatic evaluation of image similarity using different metrics, while the second is a human evaluation survey. Since the task of identifying whether two images portray the same character is subjective and difficult, we consider both qualitative and quantitative evaluation

approaches. This helps us better understand the effect of the prompt type on the quality of the different generated images. In this section, we first describe the evaluation framework used, explaining the different metrics we took into account. Then, we present the obtained results.

5.1. Evaluation Framework

5.1.1. Automatic evaluation

We utilise automated evaluation methods based on three image comparison metrics:

- *UQI* [30]: computes a pixel-based similarity score by comparing generated images with their corresponding ground-truth images. Notably, since the majority of the original images are in grayscale, the similarity computation also takes into account their grayscale versions. UQI evaluates "image quality based on factors such as loss of correlation, luminance distortion, and contrast distortion" [30].
- *CLIPscore* [31]: leverages embeddings produced by a contrastive language-image pre-trained model [5]. It is used for measuring image-caption compatibility by comparing image and text embeddings using cosine similarity. CLIP embeddings can be used for image-to-image comparisons as well, which we did by using the image encoder of the CLIP-Visual Transformer model [32]: ViT-L/14.
- *FID* [33]: is an improved version of Inception Distance (IS) proposed to measure the quality of images produced by generative models.

Since the computation of the FID metric is more time-consuming than the other two, we compute it only on a small subset of our dataset consisting of images generated for ten random characters. On the other hand, UQI and CLIPscores are computed on the entire dataset.
Additionally, we employ statistical methods for evaluating if the metrics above can measure the impact of the prompt on the quality of the generated images. For this purpose, we perform ANOVA to measure the effect of the prompt on the metric. We also perform Tukey's HSD (honestly significant difference) tests on the metrics to reflect the prompts' effect on the generated images. These statistical methods were computed on two subsets of our dataset: characters that have DBpedia abstracts and characters that do not.
Finally, we performed several Student tests to evaluate if a given property (e.g., the gender and occupation of a character) could lead to better results, and we separately made the test only on the values of the *instance of* property (P31). To carry out these tests we extract the types and properties used more than 100 times. For each property, we build two subsets. The first one includes evaluation metric results of the characters that contain the property, and the second is built by randomly choosing characters that do not have the evaluated property.

5.1.2. Human evaluation

Although the above-mentioned evaluation metrics can provide automatic measures to compare images, they are still unreliable in comparing whether the generated images successfully portray the same characters as the ground-truth images. This is because noise such as the image style or its color can affect the results of the automatic metrics. Therefore, we conduct a human evaluation study in which we ask participants to rate how likely it is that a pair of images (consisting of 1. the ground-truth image and 2. the generated image) portray the same character.

Additionally, we ask participants to list the criteria they think about when comparing two images. The latter was done to get an idea of important features to look for when generating images of fictional characters. For evaluating the agreement of the participants we compute Krippendorff's Alpha [34] on three levels: globally, per evaluated image, and per prompt type.

5.2. Evaluation Results

5.2.1. Automatic Evaluation

The results we obtained from the automatic evaluation metrics show different outcomes. In terms of UQI, all four prompt types yield a similar average similarity score of ca. 0.5, concluding that this metric is not optimal for our purposes. FID results (Table 5) show that images created from DBpedia are more similar to the ground-truth. When it comes to CLIPscores, we see a hierarchy of prompt types in terms of the obtained average similarity scores, with images generated by basic labels being the least similar (with a score of 0.48), followed by plain triples (with a score of 0.55), verbalised triples (with a score of 0.56), and lastly DBpedia abstract prompt seem to generate the most similar images to the ground-truth with a CLIPscore of 0.6. Results for UQI ad CLIP are shown in Table 4. It is important to note that contrary to UQI and CLIPscores, the FID metric is performed only on a subset of images generated for ten fictional characters, which makes it hard to make any concrete conclusions about this method.
The ANOVA conducted on the UQI and the CLIPscores is shown in Table 6 and Table 7. The results show that the UQI is not able to underline a significant difference between the prompt type as a main fixed effect on the quality of the generated image. In contrast, CLIPscores are able to reflect this effect with high confidence.
The results of Tukey's HSD test are shown in Tables 8 and 9. They highlight that the basic prompt is generally the worst prompt strategy and that the DBpedia abstract prompt is always the best one. However, for characters that do not have a DBpedia abstract, the verbalised triples prompt is better than the basic label and the plain triples prompts (with a p-value of 0.05810). Additionally, in order to understand if the number of relations and unique relation type attached to a given entity in extracted triples has an effect on the generated image quality, we compute the correlation between these variables with the CLIPscores. Results indicate that there is no such correlation (see Table 10).
Finally, we present the results of the Student tests related to the effect of the values of the *instance of* properties attached to an entity on the quality of generated images in two parts. The first displays the effect of values of the *instance of* properties on the generated images in Table 11 for plain triples prompts, and Table 13 for verbalised triples prompts. We can see that characters that already have a widely known visual representation (e.g. characters from comics, cartoons, or movies) generally have low CLIPscores. On the other hand, characters that do not have a visual representation (e.g. from written works such as novels) are usually more similar to the ground-truth images. The second part of the Student tests deals with the effect of properties on the quality of the generated images. Results are shown in Table 12 and Table 14. These results show that the majority of relations impact CLIPscores negatively.

5.2.2. Human Evaluation

To measure how humans evaluate the similarity of generated and ground-truth images, we ran an evaluation survey in which each participant is presented with images of ten different fictional characters chosen randomly from our dataset (shown in Figure 5). For each character, four pairs of images were displayed. Each pair consisted of the ground-truth image and a generated image. Participants were asked to rate how likely it is that both images portray the same character on a scale of 1-5, 1 being very unlikely and 5 being very likely. Figure 3 shows the distribution of participant replies for all ten characters based on prompt types. We immediately notice that images generated based on the three prompt types of basic labels, plain triples, and verbalised triples are more likely to be evaluated as not similar to the ground-truth image (i.e. the most frequent response for all three prompt types is one). On the other hand, images generated with DBpedia abstract prompts are most frequently rated as 3 and 4, both having the same number of responses. When examining the high numbers on the scale that indicate a high similarity between the ground truth and the generated images (i.e. 4 and 5), we notice a specific trend. The least frequent prompt type for those numbers is the basic label, followed by plain triples, verbalised triples, and DBpedia abstracts.

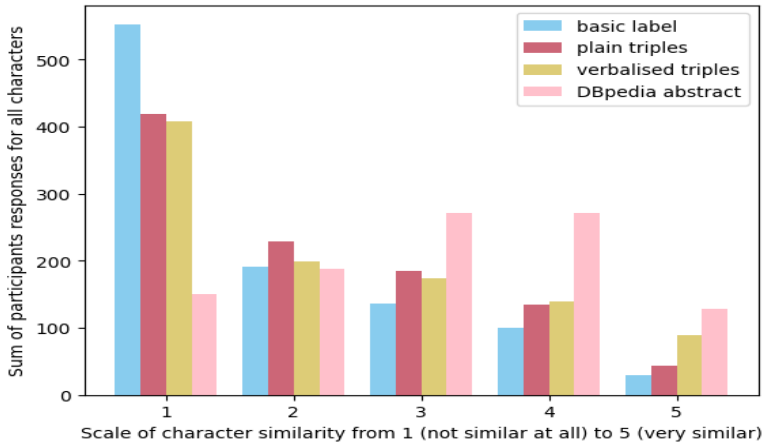

Figure 3: Distribution of the human evaluation survey results for all four prompt types.

Additionally, we analyse participant responses to the open question of which criteria they consider when giving their responses. Figure 4 presents the top ten criteria mentioned by participants. The analysis was done by extracting nouns and adjectives, and filtering out stop words and generic terms such as 'character'. We also manually grouped synonymous concepts such as *clothes*, *clothing*, and *outfit*.

In total, our survey had 101 participants ranging between the ages of 17-59 with an average age of 30. About 57% of participants were male, 41% female, and 2% non-binary. 48% of the participants had a master's education level. We did not target any specific group since we wanted to receive general responses regarding the similarity of images. Thus, we distributed the survey among friends and colleagues both from within and outside the research community. The cultural backgrounds of participants ranged from South and North America (ca. 8%) to

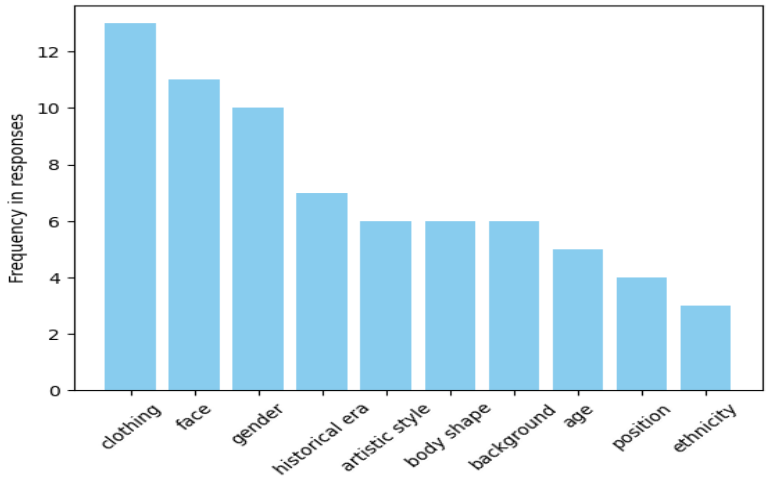

Figure 4: Results of the human evaluation survey related to the question about the key elements that have influenced user's evaluation.

Europe (ca. 63%), East Asia (ca. 10%), and the Middle East (ca. 20%). That being said, ca. 25% of participants were Italian. We received 37 responses for the open question of listing relevant criteria.
Finally, we measured the agreement of participants using Krippendorff's alpha. The global score is equal to 0.17, meaning that no concrete agreement was found. The same conclusion could be drawn at the level of the images presented to the participants and at the level of the prompt types used for the image generation (cf. Table 2 and Table 3).

5.2.3. Automatic and Human Evaluation Alignment

As a last step in our evaluation, we want to assess if there is an alignment between the score of the automatic metrics (UQI, CLIPscore, and FID) and the human evaluation. In order to be able to normalise the participant evaluations, we standardise the score given by each participant. For $i \in \{1, .., 101\}$, the unique ID of a given participant, and $j \in \{0, 39\}$, a given generated image, the standardised score is computed as follows:

$$x_{i,j}^{stand} = \frac{x_{i,j} - \mu_i}{\sigma_i}$$

The alignment between automatic metrics and human evaluation scores is shown in figure 6. We see that CLIPscores seem to be most correlated to the human scores with a Pearson correlation of 0.5 for the plain triples prompt, 0.6 for the verbalised triples prompt, and 0.7 for the basic label prompt. Concerning the DBpedia abstract prompt, none of the metrics seem to be correlated with the human evaluation. UQI and FID are not correlated to human evaluation, results, both having scores close to zero.

6. Discussion

Results of most automatic evaluation approaches we used (CLIPscores, ANOVA, and Tukey's HSD) as well as the human evaluation results suggest a clear trend: images generated using DBpedia abstracts as prompts were rated as most similar to the ground truth images, followed respectively by verbalised triples prompts, plain triples prompts, and basic label prompts. This implies that DBpedia abstracts, which are written by human editors and contain more natural, diverse, and fluent text, enable text-to-image generators to produce better results. The fact that verbalised triple prompts produce the second-best results further emphasises the importance of fluent text on the quality of the generated image. These results directly answer **RQ1**.
When further analysing the obtained CLIPscores, we observe that the maximum CLIPscore occurred for an image generated by using a basic label prompt, possibly indicating that the text-to-image model had "seen" this character during training. This enabled it to create a similar image to the ground-truth one without adding any additional context. However, the lowest CLIPscore also occurred when using the prompt type of basic label, further emphasising that for some characters, generating an image based only on their label is not enough.
We conclude that in order to automatically generate images for fictional characters that correctly portray them, using natural text descriptions is the best option. When this text is available (e.g. in DBpedia abstracts), it is best to use it, however, as we have observed when creating our dataset, many entities of fictional characters (See Table 1) do not have a DBpedia abstract. To create images for those characters, the best method seems to be extracting knowledge about them in the form of triples, verbalising those triples using a large language model, and giving the verbalised text as input to a text-to-image generative model. In this case, the content of the triples is crucial for generating high-quality images. However, the quality of images is not related to the number of triples or the number of unique relations contained in the triples. But is highly dependent on object values, highlighting the impact of the value of *instance of* property of fictional characters on the quality of generated images. Answering **RQ2**, generated images can then be leveraged for completing missing images in Wikidata entities.
Finally, addressing **RQ3**, when comparing the three automatic evaluation metrics we see that only the CLIPscores align with the human evaluation scores. This is because, unlike the human and CLIP evaluations which assess semantic similarity, the UQI and FID metrics only focus on image quality. This limitation in evaluating semantic content likely explains the discrepancy in results between the three automatic evaluation metrics.

7. Limitations and Risks

Our work is limited in many aspects. First, we are currently dealing only with English data due to the limitations of the verbalisation model and the Stable Diffusion model we used. Future work will consider dealing with multilingual datasets as well.
Additionally, when designing prompts based on Wikidata triples, we had to make decisions such as extracting triples based on subjects without considering objects. We also treated all triples equally with no emphasis on properties or types of entities. As shown by the open question in our human evaluation survey, it is evident that some properties are more important than others

when generating images to portray a specific character. Future work can potentially explore in more depth which properties lead to better representations of characters. Further, when encountered with triples that have the same predicate, we selected the one with the longest object assuming it would contain more information. We are aware that this decision might have removed important information for characters, and this can be addressed in future work by concatenating object strings or summarising them automatically.

Our usage of the Stable Diffusion generative model means that our method is inheriting its biases as well. Although directly leveraging information about each character from its triples is supposed to limit biases when generating images, this cannot always be controlled (e.g., for some female entities, the model generated images of male characters). Additionally, using a pre-defined set of negative prompts for all characters (which includes terms such as *mutilated* and *disfigured*) is a considerable limitation of the model to correctly portray characters. A possible solution for this could be to design specific negative prompts for each individual character in a semi-automatic manner or to use another type of text-to-image model that does not require negative prompting.

Our work is also limited in terms of the ground-truth dataset constructed based on Wikidata entities that already have images. This is because, for some of these entities, the images are not reliably portraying the character, but the actor depicting the character.

Finally, in order to mitigate any copyright and/or privacy risks, we stress that our method is not suggested to be directly deployed into Wikidata, as we think that using AI-generated images can potentially be very harmful. Should this method be used for image completion, we encourage clearly watermarking images as AI-generated.

8. Conclusion

In this paper, we investigate four different methods for generating prompts based on extracting knowledge in the form of triples. We then generate images based on each prompt using Stable Diffusion, a generative text-to-image model. We evaluate the different prompt types by automatic as well as human evaluation approaches and conclude that the best-generated images are based on natural language text that includes the context and background of the character. When possible, this text can be extracted from a human-edited source such as DBpedia abstracts, however, most characters do not have a DBpedia entity. This brings to light the need to verbalise triples (i.e. transform them into natural text based on large language models) and use them as prompts in order to receive the best visual representation of their corresponding fictional characters. To the best of our knowledge, our work is novel in terms of utilising triples for prompt engineering in order to complete missing information on Wikidata. Possible future work includes finetuning the last Stable Diffusion model via a Lora adaptation [35], trying other text-to-image models that rely on different architectures, and modifying prompts to include the most significant triples by investigating which properties affect image quality the most. Our approach is not intended to directly complete entities on Wikidata with AI-generated images, rather it can be used by editors to further enrich entities such as fictional characters, fictional places, or landscapes. Alternatively, instead of directly using the output of generative models, they could be given to artists who can use them as inspiration to create depictions of entities.

Acknowledgments

This project is the result of a research task force team at the International Semantic Web Summer School (ISWS) 2023. We would like to thank the organisers and tutors. We especially thank our mentor Albert Meroño-Peñuela for his valuable advice and input throughout this work.
We made use of the central High Performance Computer system at Freie Universität Berlin to conduct the data collection and image generation parts, and we would like to express our gratitude for the resources provided.
The work of the author, Sefika Efeoglu, is funded by the German Federal Ministry of Education and Research (BMBF) and the state of Berlin under the Excellence Strategy of the Federal Government and the Länder over the project.
This paper has been developed within the HE project MuseIT, which has been cofounded by the European Union under the Grant Agreement No 101061441. Views and opinions expressed are, however, those of the authors and do not necessarily reflect those of the European Union or European Research Executive Agency.
This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] D. Vrandečić, Wikidata: A New Platform for Collaborative Data Collection, in: Proceedings of the 21st international conference on world wide web, 2012, pp. 1063–1064.
[2] D. Abián, A. Meroño-Peñuela, E. Simperl, An Analysis of Content Gaps Versus User Needs in the Wikidata Knowledge Graph, in: International Semantic Web Conference, Springer, 2022, pp. 354–374.
[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis With Latent Diffusion Models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
[4] Introducing ChatGPT, 2023. URL: https://openai.com/blog/chatgpt.
[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.
[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, ArXiv abs/2204.06125 (2022).
[7] F. Schneider, Z. Jin, B. Schölkopf, Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion, arXiv preprint arXiv:2301.11757 (2023).
[8] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, Text-to-image Diffusion Models in Generative AI: A Survey, arXiv preprint arXiv:2303.07909 (2023).
[9] E. Mansimov, E. Parisotto, J. L. Ba, R. Salakhutdinov, Generating Images from Captions with Attention, arXiv preprint arXiv:1511.02793 (2015).
[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative Adversarial Text to Image Synthesis, in: International conference on machine learning, PMLR, 2016, pp. 1060–1069.
[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907–5915.
[12] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316–1324.
[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.
[14] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al., Cogview: Mastering text-to-image generation via transformers, Advances in Neural Information Processing Systems 34 (2021) 19822–19835.
[15] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, Nüwa: Visual Synthesis Pre-training for Neural visUal World creAtion, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI, Springer, 2022, pp. 720–736.
[16] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes, CoRR abs/1312.6114 (2013).
[17] A. Vahdat, J. Kautz, NVAE: A Deep Hierarchical Variational Autoencoder, 2021. arXiv:2007.03898.
[18] R. Child, Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images, ArXiv abs/2011.10650 (2020).
[19] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, arXiv preprint arXiv:2112.10741 (2021).
[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, Advances in Neural Information Processing Systems 35 (2022) 36479–36494.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
[22] V. Liu, L. B. Chilton, Design Guidelines for Prompt Engineering Text-to-Image Generative Models, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–23.
[23] J. Oppenlaender, R. Linder, J. Silvennoinen, Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering, arXiv preprint arXiv:2303.13534 (2023).
[24] J. Oppenlaender, A Taxonomy of Prompt Modifiers for Text-to-Image Generation, arXiv preprint arXiv:2204.13988 (2022).
[25] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, et al., Prompt Engineering for Healthcare: Methodologies and Applications, arXiv preprint arXiv:2304.14670 (2023).
[26] G. Amaral, O. Rodrigues, E. Simperl, WDV: A Broad Data Verbalisation Dataset Built from Wikidata, in: U. Sattler, A. Hogan, M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web – ISWC 2022, Springer International Publishing, Cham, 2022, pp. 556–574.
[27] L. F. R. Ribeiro, M. Schmitt, H. Schütze, I. Gurevych, Investigating Pretrained Language Models for Graph-to-Text Generation, in: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, Association for Computational Linguistics, Online, 2021, pp. 211–227. URL: https://aclanthology.org/2021.nlp4convai-1.20. doi:10.18653/v1/2021.nlp4convai-1.20.
[28] Brümmer, Martin and Dojchinovski, Milan and Hellmann, Sebastian, DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 3339–3343. URL: https://aclanthology.org/L16-1532.
[29] Y. Chen, X-IQE: eXplainable Image Quality Evaluation for Text-to-Image Generation with Visual Large Language Models, arXiv preprint arXiv:2305.10843 (2023).
[30] D. Varga, Full-Reference Image Quality Assessment Based on an Optimal Linear Combination of Quality Measures Selected by Simulated Annealing, Journal of Imaging 8 (2022). URL: https://www.mdpi.com/2313-433X/8/8/224. doi:10.3390/jimaging8080224.
[31] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. URL: https://aclanthology.org/2021.emnlp-main.595. doi:10.18653/v1/2021.emnlp-main.595.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.
[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, 2018. arXiv:1706.08500.
[34] A. F. Hayes, K. Krippendorff, Answering the Call for a Standard Reliability Measure for Coding Data, Communication Methods and Measures 1 (2007) 77–89. URL: https://doi.org/10.1080/19312450709336664. doi:10.1080/19312450709336664. arXiv:https://doi.org/10.1080/19312450709336664.
[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, 2021. arXiv:2106.09685.

A. Appendix

Table 2
Krippendorf's alpha focused on per character agreement

Image	Krippendorf alpha
Lancelot (Q215681)	0.09354
Fëanor (Q716794)	0.09049
John Sheppard (Q923684)	0.00390
Hoshi Sato (Q1055776)	0.03129
Puck (Q1248616)	0.04627
Harry Potter (Q3244512)	0.08192
Agramante (Q3606846)	0.08182
Mariner Moose (Q5353616)	0.08347
Octobriana (Q7077012)	0.04676
Phanuel (Q7180638)	0.13665

Table 3
Krippendorf's alpha focused on per prompt type agreement

Prompt	Krippendorf alpha
Basic Label	0.17647
Plain Triples	0.12030
Verbalised Triples	0.20725
DBpedia Abstract	0.03761

Table 4
Comparison between generated and ground-truth images based on UQI and CLIP

Prompt Type	min UQI	mean UQI	max UQI	min ClipSim	mean ClipSim	max ClipSim
Basic Label	0.0	0.49970	0.80195	0.14747	0.48277	**0.9590**
Plain Triples	0.0	0.49966	**0.87212**	**0.18396**	0.54535	0.88957
Verbalised Triples	0.0	**0.50075**	0.86605	0.17431	0.55710	0.92599
DBpedia Abstract	0.0	0.49151	0.78226	0.16856	**0.60192**	0.92902

Table 5
FID metrics computed on the human evaluation subset (note that lower numbers mean higher similarity)

Fictionnal Character	Basic label	Plain prompt	Verbalised prompt	DBpedia abstract
Lancelot (Q215681)	126.35087	**54.766038**	96.49390	67.89936
Fëanor (Q716794)	119.39887	**118.62600**	125.80051	123.55941
John Sheppard (Q923684)	241.30466	236.92710	274.41479	**139.81330**
Hoshi Sato (Q1055776)	228.71975	203.96307	225.44356	**161.90406**
Puck (Q1248616)	**118.86526**	157.89964	137.24375	145.40420
Harry Potter (Q3244512)	190.61544	217.70724	197.86598	**188.29662**
Agramante (Q3606846)	125.21297	**67.034159**	132.19362	110.24826
Mariner Moose (Q5353616)	**73.558398**	174.05515	104.00699	76.30708
Octobriana (Q7077012)	**78.224097**	209.75924	173.35200	204.09481
Phanuel (Q7180638)	164.80482	75.619137	**49.82890**	57.28145
Average	146.70551	151.63568	151.66440	**127.48086**

Table 6
Analysis of variance focused on entity with abstracts (N=914) regarding the distribution of UQI and the CLIPscore, with the prompt strategy as the main fixed effects.

Metric	df	sum of squares	mean of squares	F value	significance
CLIPscore	3	4.77477	1.59159	100.29899	**2.18256e-62**
UQI	3	0.02104	0.00701	0.53132	0.66078

Table 7
Analysis of variance focused on entity without abstracts (N=586) regarding the distribution of UQI and the CLIPscore, with the prompt strategy as the main fixed effects.

Metric	df	sum of squares	mean of squares	F value	significance
CLIPscore	2	2.29885	1.1494	77.94544	**4.47908e-33**
UQI	2	0.022125	0.00737	0.51729	0.59622

Table 8
Pairwise tests using Tukey HSD related to the effect of the prompt on the CLIPscore, on the subset of images having DBpedia abstracts

prompt1	prompt2	diff	lower	upper	q-value	p-value
basic prompt	plain prompt	0.05848	0.04334	0.07363	14.03695	0.001
basic prompt	verbalised prompt	0.06724	0.05209	0.08238	16.13806	0.001
basic prompt	dbpedia abstract prompt	0.10023	0.08508	0.11537	24.05521	0.001
plain prompt	verbalised prompt	0.00875	-0.00639	0.02390	2.10110	0.44767
plain prompt	dbpedia abstract prompt	0.04174	0.02659	0.05688	10.01826	0.001
verbalised prompt	dbpedia abstract prompt	0.03298	0.01784	0.04813	7.91715	0.001

Table 9
Pairwise tests using Tukey HSD related to the effect of the prompt on the CLIPscore, on the subset of images do not having DBpedia abstracts

prompt1	prompt2	diff	lower	upper	q-value	p-value
basic prompt	plain prompt	0.06934	0.05220	0.08649	13.41698	0.001
basic prompt	verbalised prompt	0.08605	0.06890	0.10320	16.6495	0.001
plain prompt	verbalised prompt	0.01670	-0.00043	0.03385	3.23256	0.05810

Table 10
Correlation scores between number of relation and CLIP score for triple-based prompts

var	plain triple	verbalised triples
CLIP score Vs. Number of relations	0.10433	0.15312
CLIP score Vs. Number of unique relations	0.16794	0.17828

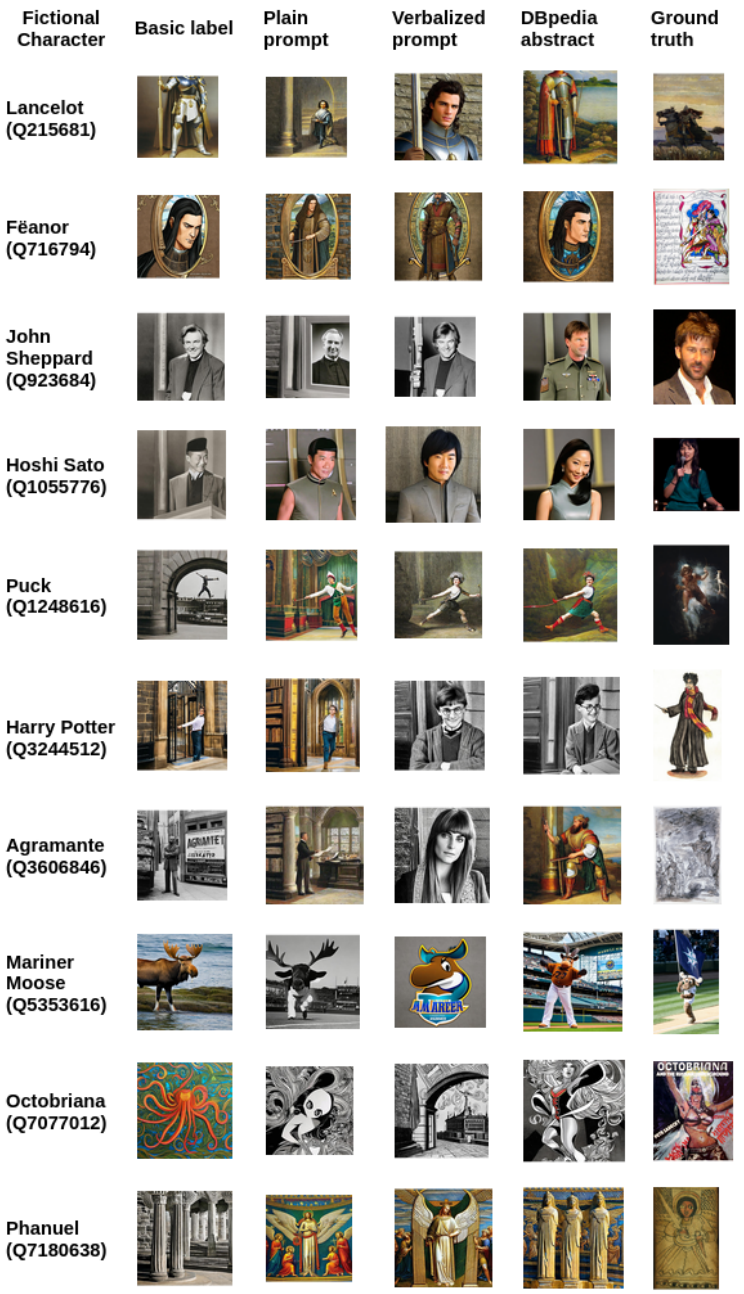

Figure 5: The 10 random generated images used for the Human Evaluation

Table 11
Student tests on plain triples on the values of the "instance of" relations appearing more than 100 times

instance of	sample size	mean with	mean without	t student	p-value
graphic novel character	120	0.53759	0.60367	-4.85939	**0.0**
fictional character in comics	120	0.54465	0.60367	-4.14029	**5e-05**
comic character	120	0.54554	0.60367	-4.12248	**5e-05**
comics characters	120	0.54714	0.60367	-3.80711	**0.00018**
comic strip character	120	0.54894	0.60367	-3.78012	**0.0002**
cartoon character	151	0.54186	0.59345	-3.66697	**0.00029**
comic characters	120	0.55257	0.60367	-3.53789	**0.00048**
fictional man	604	0.56867	0.54659	3.25443	**0.00117**
fictional character appearing in a film	146	0.5415	0.58396	-3.05048	**0.0025**
human being that only exists in fictional works	604	0.56749	0.54659	2.99562	**0.00279**
human fictional character	604	0.5668	0.54659	2.95531	**0.00318**
comic book character	120	0.5613	0.60367	-2.90949	**0.00396**
fictional person	604	0.56611	0.54659	2.86978	**0.00418**
character in a book	311	0.57115	0.54416	2.81644	**0.00501**
animation character	151	0.55279	0.59345	-2.77323	**0.0059**
fictional persons	604	0.56374	0.54659	2.50207	**0.01248**
fictional character who appears in animated films, television, and other animated works	151	0.55864	0.59345	-2.43599	**0.01543**
television show character	207	0.54562	0.57471	-2.42813	**0.0156**
fictional woman	604	0.5623	0.54659	2.27222	**0.02325**
fictional character who appears in a television series	207	0.54721	0.57471	-2.27788	**0.02325**
human fictional characters	604	0.56201	0.54659	2.26597	**0.02363**
comics character	120	0.57095	0.60367	-2.22542	**0.02699**
fictional human	604	0.56125	0.54659	2.16103	**0.03089**
animated character	151	0.56282	0.59345	-2.16122	**0.03147**
cartoon characters	151	0.5668	0.59345	-1.91238	0.05678
TV show character	207	0.55195	0.57471	-1.8769	0.06124
TV character	207	0.55358	0.57471	-1.77022	0.07743
cinematic character	146	0.55876	0.58396	-1.7426	0.08246
fictional character appearing in written works	311	0.56126	0.54416	1.72544	0.08495
character in literature	311	0.56107	0.54416	1.71958	0.08601
movie character	146	0.55973	0.58396	-1.69479	0.09119
book character	311	0.56006	0.54416	1.63574	0.1024
literary character	311	0.55985	0.54416	1.62197	0.10532
novel character	311	0.55851	0.54416	1.43847	0.15081
literature character	311	0.55855	0.54416	1.42897	0.15352
film character	146	0.56439	0.58396	-1.38642	0.16668
TV series character	207	0.55938	0.57471	-1.25831	0.20899
television series character	207	0.56432	0.57471	-0.90818	0.36431
character in a novel	311	0.553	0.54416	0.90722	0.36464
human (as opposed to supernatural) character in the Old Testament/Hebrew Bible or New Testament	112	0.53849	0.55244	-0.85578	0.39304
human biblical figure	112	0.54104	0.55244	-0.7692	0.44259
television character	207	0.56607	0.57471	-0.75946	0.44801
biblical human	112	0.54532	0.55244	-0.47913	0.63232
biblical human character	112	0.54976	0.55244	-0.17741	0.85935
human in the Bible	112	0.55174	0.55244	-0.04366	0.96521

Table 12
Student tests on plain triples on the relations appearing more than 100 times

relation	sample size	mean with	mean without	t student	p-value
said to be the same as	134	0.52624	0.59053	-4.45282	**1e-05**
described by source	200	0.54037	0.58667	-3.94038	**0001**
different from	235	0.53497	0.57629	-3.85235	**00013**
topic's main category	135	0.54547	0.60322	-3.88675	**00013**
father	216	0.53967	0.57929	-3.56135	**00041**
from narrative universe	392	0.53984	0.56521	-37027	**00221**
place of birth	169	0.53054	0.57046	-3382	**00257**
sibling	157	0.53544	0.57294	-2.73217	**00665**
name in native language	155	0.54605	0.57979	-2.64531	**00858**
given name	493	0.55413	0.53382	2.58674	**00983**
enemy	176	0.55583	0.58792	-2.53043	**01183**
child	181	0.54056	0.57175	-2.52187	**0121**
part of	133	0.54286	0.57954	-2.43042	**01575**
mother	144	0.55002	0.58305	-2.40593	**01677**
media franchise	145	0.54594	0.57799	-2.2482	**02532**
first appearance	146	0.54038	0.56965	-2.15909	**03166**
present in work	389	0.5286	0.54332	-1.69822	08987
country of citizenship	428	0.55127	0.53896	1.49145	0.13621
languages spoken, written or signed	293	0.54282	0.55741	-1.45697	0.14566
family name	281	0.55035	0.53763	1.23692	0.21664
spouse	215	0.54113	0.55489	-1.1824	0.2377
member of	136	0.55572	0.57103	-1.10956	0.26817
narrative role	177	0.54159	0.55615	-19498	0.27427
residence	143	0.54038	0.55417	-0.95292	0.34144
occupation	565	0.5409	0.54756	-0.92396	0.3557
creator	588	0.54809	0.54479	0.46119	0.64475
voice actor	110	0.53768	0.54433	-0.4121	0.68067
eye color	113	0.54737	0.55021	-0.17031	0.86492
sex or gender	286	0.55365	0.55528	-0.16341	0.87025
hair color	114	0.55313	0.55122	0.11972	0.90481
performer	438	0.54877	0.54901	-02778	0.97785

Table 13
Student tests on verbalised triples on the values of the "instance of" relations appearing more than 100 times

instance of	sample size	mean with	mean without	t student	p-value
fictional person	604	0.55721	0.52212	52155	**0**
human being that only exists in fictional works	604	0.55742	0.52212	58938	**0**
fictional man	604	0.56348	0.52212	5.93812	**0**
fictional persons	604	0.5636	0.52212	62202	**0**
fictional woman	604	0.56756	0.52212	6.67694	**0**
human fictional character	604	0.56339	0.52212	5.96221	**0**
human fictional characters	604	0.56621	0.52212	6.46184	**0**
fictional human	604	0.56144	0.52212	5.6847	**0**
human (as opposed to supernatural) character in the Old Testament/Hebrew Bible or New Testament	112	0.51176	0.59839	-5.70389	**0**
comics character	120	0.52673	0.58356	-4.15193	**5e-05**
human biblical figure	112	0.5395	0.59839	-3.79607	**00019**
fictional character who appears in animated films, television, and other animated works	151	0.52863	0.58071	-3.76422	**0002**
fictional character in comics	120	0.5314	0.58356	-3.77533	**0002**
fictional character appearing in written works	311	0.56562	0.52885	3.72153	**00022**
biblical human character	112	0.54161	0.59839	-3.57517	**00043**
comic book character	120	0.5355	0.58356	-3.41365	**00075**
graphic novel character	120	0.53706	0.58356	-3.2513	**00132**
comic characters	120	0.54027	0.58356	-3.24203	**00136**
human in the Bible	112	0.55056	0.59839	-3.18845	**00164**
cartoon character	151	0.53912	0.58071	-38265	**00224**
comic character	120	0.54249	0.58356	-2.9583	**00341**
biblical human	112	0.55704	0.59839	-2.80093	**00555**
comic strip character	120	0.54284	0.58356	-2.76604	**00612**
character in a novel	311	0.553	0.52885	2.53108	**01162**
character in literature	311	0.5534	0.52885	2.5077	**01241**
cartoon characters	151	0.54863	0.58071	-2.43456	**01549**
novel character	311	0.55218	0.52885	2.37615	**0178**
comics characters	120	0.54978	0.58356	-2.32572	**02087**
animated character	151	0.54876	0.58071	-2.26202	**02441**
cinematic character	146	0.5397	0.57174	-2.25485	**02489**
literature character	311	0.55055	0.52885	2.19677	**02841**
movie character	146	0.53912	0.57174	-2.1486	**03249**
literary character	311	0.54767	0.52885	1.89849	0581
TV character	207	0.56049	0.53791	1.8604	06354
character in a book	311	0.54643	0.52885	1.76267	07845
fictional character appearing in a film	146	0.54601	0.57174	-1.72412	08575
book character	311	0.54493	0.52885	1.63332	0.10291
film character	146	0.54905	0.57174	-1.56495	0.11868
fictional character who appears in a television series	207	0.55602	0.53791	1.47277	0.14158
animation character	151	0.56582	0.58071	-1.1429	0.25399
television character	207	0.55208	0.53791	1.14182	0.25419
TV show character	207	0.52439	0.53791	-16176	0.28896
television show character	207	0.55049	0.53791	0.99373	0.32094
TV series character	207	0.54779	0.53791	0.78829	0.43098
television series character	207	0.54323	0.53791	0.44074	0.65964

Table 14
Student tests on verbalised triples on the relations appearing more than 100 times

relation	sample size	mean with	mean without	t student	p-value
from narrative universe	392	0.54247	0.59269	-6.21662	**0**
enemy	176	0.55843	0.60108	-3.18641	**00157**
eye color	113	0.54021	0.59086	-3.12825	**00199**
father	216	0.55582	0.58778	-2.84772	**00461**
media franchise	145	0.56023	0.5985	-2.74637	**00641**
present in work	389	0.53543	0.55898	-2.72472	**00658**
topic's main category	135	0.55616	0.59396	-2.55165	**01128**
member of	136	0.57542	0.60546	-2.36973	**0185**
languages spoken, written or signed	293	0.55711	0.57953	-2.29727	**02196**
name in native language	155	0.56111	0.58959	-2.16698	**031**
mother	144	0.56521	0.59498	-2.16136	**0315**
sibling	157	0.54875	0.57697	-2.15316	**03207**
first appearance	146	0.55121	0.58171	-2.12306	**0346**
part of	133	0.55823	0.58861	-2.11582	**0353**
occupation	565	0.54869	0.56324	-26606	**03905**
said to be the same as	134	0.55461	0.5818	-1.91761	05623
place of birth	169	0.55391	0.57885	-1.90071	0582
voice actor	110	0.54717	0.57493	-1.7728	07766
different from	235	0.54767	0.56429	-1.60514	0.10914
hair color	114	0.56243	0.58611	-1.56321	0.1194
given name	493	0.56572	0.55461	1.44034	0.15009
performer	438	0.55708	0.56849	-1.37983	0.16799
child	181	0.54882	0.56591	-1.33046	0.18421
family name	281	0.55684	0.56664	-11722	0.30949
narrative role	177	0.55689	0.5697	-0.97679	0.32934
sex or gender	286	0.55629	0.56362	-0.71496	0.47493
creator	588	0.5554	0.55862	-0.45502	0.64918
instance of*	27	0.58846	0.57998	0.25397	0.80052
residence	143	0.55984	0.5634	-0.25201	0.80121
spouse	215	0.56128	0.56394	-0.23058	0.81775
described by source	200	0.56531	0.56379	0.13335	0.89399
country of citizenship	428	0.55895	0.55945	-06071	0.95161

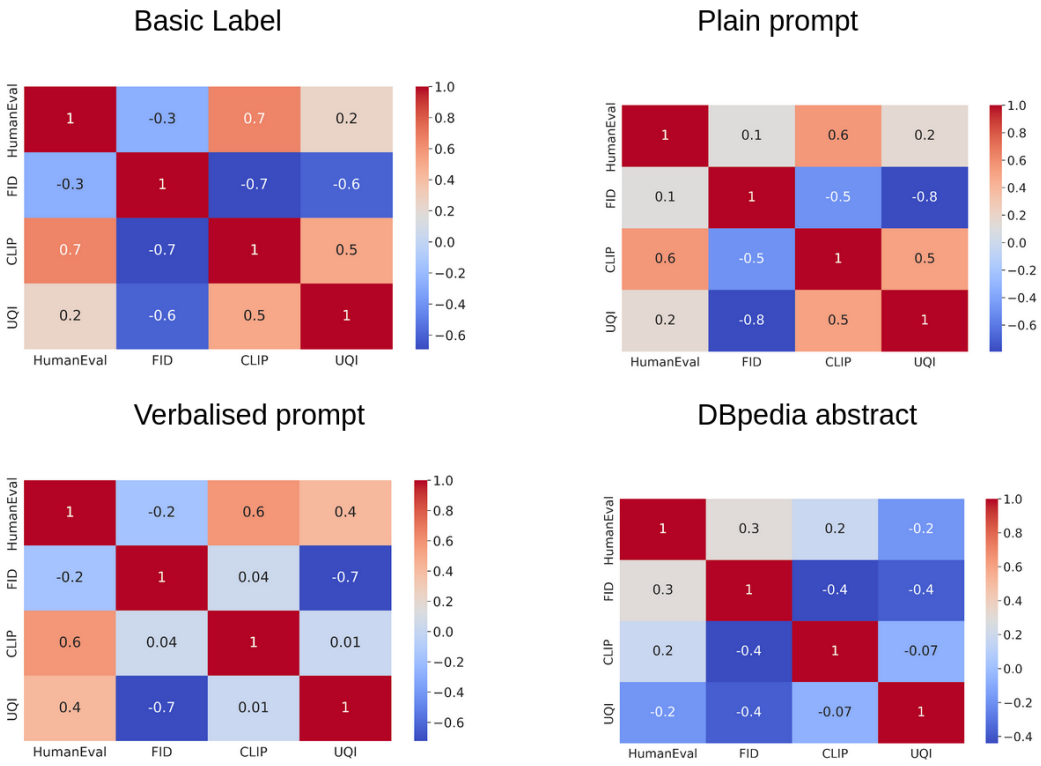

Figure 6: Correlation plots of the automatic and the human evaluation

Table 15
Prompt examples of an entity from the generated prompt dataset.

Fictional Character	Lancelot (Q215681)
Basic Label	Lance Hunter
Plain Triples	Lance Hunter instance of fictional character in comics. Lance Hunter instance of comics character. Lance Hunter instance of comic book character. Lance Hunter instance of comic character. Lance Hunter instance of comic strip character. Lance Hunter instance of comics characters. Lance Hunter instance of comic characters. Lance Hunter instance of graphic novel character. Lance Hunter instance of human being that only exists in fictional works. Lance Hunter instance of fictional human. Lance Hunter instance of fictional person. Lance Hunter instance of fictional man. Lance Hunter instance of fictional persons. Lance Hunter instance of fictional woman. Lance Hunter instance of human fictional character. Lance Hunter instance of human fictional characters. Lance Hunter instance of TV show character. Lance Hunter instance of TV character. Lance Hunter instance of television series character. Lance Hunter instance of television show character. Lance Hunter instance of TV series character. Lance Hunter instance of fictional character who appears in a television series. Lance Hunter instance of television character. Lance Hunter present in work Marvel's Agents of S.H.I.E.L.D.. Lance Hunter from narrative universe shared fictional universe of many comic books published by Marvel Comics. Lance Hunter given name male given name. Lance Hunter sex or gender to be used in sex or gender (P21) to indicate that the human subject is a male or semantic gender (P10339) to indicate that a word refers to a male person. Lance Hunter family name Hunter family name.
Verbalised Triples	Lance Hunter is a fictional character in Marvel's Agents of S.H.I.E.L.D. He is a character in the TV series Marvel's Agents of S.H.I.L.D. He is also a character in the comic book genre. He is also a character in the graphic novel genre.
DBpedia Abstract	Lancelot Lance Hunter is a fictional character appearing in American comic books published by Marvel Comics. He first appeared in Captain Britain Weekly 19 (February 16, 1977) and was created by writer Gary Friedrich and artist Herb Trimpe. Hunter is a Royal Navy Commander who became Director of S.T.R.I.K.E. before later gaining the rank of Commodore and becoming Joint Intelligence Committee Chair. The character made his live-action debut in the Marvel Cinematic Universe television series Agents of S.H.I.E.L.D., portrayed by Nick Blood.