PictureStories: Predicting the Task Adherence of Language Learner Answers to a Picture-Story-Based Writing Task

Marie Bexte[1], Andrew Caines[2], Diane Nicholls[3], Paula Buttery[2], Torsten Zesch[1]
[1]CATALPA – Center of Advanced Technology for Assisted Learning and Predictive Analytics,
FernUniversität in Hagen, Germany
[2]ALTA Institute & Computer Laboratory, University of Cambridge, United Kingdom
[3]Cambridge University Press & Assessment, United Kingdom
Correspondence: marie.bexte@fernuni-hagen.de

Abstract

We investigate the automated evaluation of English language learner answers to writing tasks featuring picture stories. This is usually limited to language proficiency only, neglecting the context of the picture. Instead, our analysis focuses on task adherence, which for example allows detection of off-topic answers. Since there is a lack of suitable training and evaluation data, our first step is to build the PictureStories dataset. To this end, we develop a marking rubric that covers task adherence with respect to both form and content. Six annotators mark 713 learner answers written in response to one of five picture stories. Having assembled the dataset, we then explore to what extent task adherence can be predicted automatically. Our experiments assume a scenario where no or just a few labelled answers are available for the picture story which is being marked. For form-focused criteria, we find that it is beneficial to finetune models across tasks. With content-focused criteria, few-shot prompting Qwen emerges as the best-performing method. We examine the trade-off between including the story image vs. example answers in the prompt and find that examples suffice in many cases. While for some LLMs, few-shot prompting results may look promising on the surface, we demonstrate that a much simpler method can do just as well when shown the same examples.

1 Introduction

Production is a key part of language learning, and pictures are elegant cues to test it (Boers, 2018). Their visual nature gives them a language-independent quality, especially when they are entirely free of text. This removes any priming effect due to text in the target language.

Figure 1 shows an example: A story is illustrated in three pictures. Irrespective of their first language, learners can grasp its content and produce their version of the story in the target language.

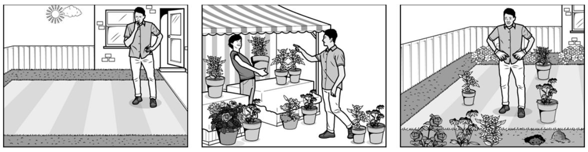


Figure 1: An example picture story from the Write & Improve platform. This *gardening* story is one of the five tasks included in PictureStories.

Many language learning tasks with a pictorial stimulus judge language quality, e.g. by evaluating spelling (Laarmann-Quante et al., 2019), assigning a Common European Framework of Reference for Languages (CEFR) level (Cambridge Assessment, 2020), or analysing syntactic structures (Köhn and Köhn, 2018). We are also interested in answer form, but specifically with respect to *task adherence*. For this, we need to go beyond form and also cover *content*, partially for student feedback but mainly as a form of cheating prevention: A well-formed, highly proficient answer should not get a perfect score if it is entirely off-topic (as the learner may have memorized it (Samant et al., 2025) or copied it from somewhere instead of producing it spontaneously in response to the stimulus).

We operationalise task adherence with a 5-item marking rubric and use it to annotate 713 answers written by language learners to describe one of five picture stories. The learners are users of the essay-writing practice platform Write & Improve[1], a free tool for language learners that is offered by Cambridge University Press & Assessment (CUP&A). We run scoring experiments on the PictureStories dataset we collect to test how well the different criteria can be predicted automatically. We assess a variety of baselines and compare their performance to score prediction with a set of open source vision-and-language LLMs.

[1] https://writeandimprove.com

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2398–2415
March 24-29, 2026 ©2026 Association for Computational Linguistics

We find that for criteria that are independent of the content shown in the story image, finetuned cross-story models do best. Where the context of the story image does matter, LLMs are superior. There is a positive effect of including the story images in the prompt, pushing performance of the best LLM to a robust level. However, this zero-shot prompting without example answers in many cases does just as well as few-shot prompting, even without inclusion of the story image.

Beyond this, our contributions are:

- We collect the PictureStories dataset, which consists of 713 learner answers that describe one of five picture stories.
- All learner answers are reliably marked by six expert raters for task adherence according to our five-item rubric.
- The dataset[2] as well as our experiment code[3] are made available for non-commercial use.

2 Related Work

The detection of off-topic answers has received some attention in purely text-based essay scoring. There are approaches both for the assignment of relevance scores to individual sentences (Higgins et al., 2006) or entire essays (Persing and Ng, 2014; Cummins et al., 2016). Some interpret relevance as a binary variable, others distinguish more than two levels of relevance. More recently, embedding spaces are being used to detect topical outliers (Huang et al., 2023; Albatarni et al., 2024).

Our study differs from the aforementioned ones, as the task is based on a picture story. We also strive for a more fine-grained assessment of task adherence, rather than just topical answer relevance. In the following, we describe previous work on the use of pictures for language learning tasks, especially with respect to automated evaluation. We then give an overview of models with vision-and-language capabilities.

2.1 Evaluating Picture-Based Language Learning Tasks

While there are corpora that do include pictures, and even picture stories, as their tasks, the pictures are often more of a means to constrain answer content (Kotani et al., 2011; Laarmann-Quante et al., 2019; Wottawa and Adda-Decker, 2016; Köhn and Köhn, 2018). These corpora are annotated for target hypotheses or used to determine language quality in general. They do not account for the alignment of learner answers with the images that they are responses to.

One exception to this is the work by King and Dickinson (2016, 2018). Their SAILS corpus contains one-sentence descriptions of cartoon images. Answers are annotated along a rubric that covers some aspects of task adherence, such as answer coverage of key elements, or whether an answer is even an attempt to fulfil the task. King and Dickinson (2018) also consider automated evaluation of picture descriptions, but approach this by using a pool of answers as the background corpus, i.e. not through inclusion of the images themselves.

Similarly, there is other work that scores language learner responses to picture stimuli, but includes the content dimension via a textual background corpus that represents the image (Somasundaran and Chodorow, 2014; Somasundaran et al., 2015) or through an elaborate rubric that spells out the different content elements that should be addressed (Baumann et al., 2024).

Rei (2017) is the exception to this and trains a model that compares photographs to learner descriptions of them. As negative examples, i.e. instances where photograph and description do not match, descriptions are paired with unrelated photographs. This take on the task is somewhat similar to the SNLI-VE dataset (Xie et al., 2019), where entailment between photographs and descriptive sentences is evaluated. Other work with a similar motivation as ours is Tanaka et al. (2022), who collect a dataset of one-sentence learner descriptions of photographs. It is annotated with corrections that also fix content errors, but these are not distinguished from purely grammatical errors.

A notable difference between the previously described work and our is that we are not working with photographs. PictureStories also has challenging negative examples that were actually written in response to a certain picture story, but were found to fall short of a criterion during manual evaluation.

2.2 Models

In recent times, a variety of models that are capable of combined processing of image and text have emerged. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are two examples that embed images and text into a shared space. This architecture

[2] https://englishlanguageitutoring.com/datasets/picture-stories

[3] https://github.com/mariebexte/picture-stories

allows for the two modalities to be compared via cosine similarity and as such facilitates zero-shot image classification.

With the advent of LLMs, there is a continuous stream of new models. Increasingly, LLMs also come with the capability of processing visual content. Examples are Qwen (Bai et al., 2023), LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Idefics (Laurençon et al., 2023), vision-capable versions of Llama (Touvron et al., 2023) and ChatGPT 4o (OpenAI, 2024).

Parallel to the development of different models, benchmarks have been created to measure and track their performance over time. In the realm of visio-linguistic educational applications, EXAMS-V is a benchmark that consists of multiple choice exam questions with visual content (Das et al., 2024). Regarding picture stories, Woloszyn and Gagl (2025) test whether GPT-4 can describe picture stories the way children do. With respect to the evaluation of learner language, Pack et al. (2024) find promising results for essay scoring. However, Benedetto et al. (2024)'s results for essay marking with LLMs are mixed: while GPT-4o-mini performs almost as well as a feature-based model, other LLMs perform poorly, even pathologically predicting the same grade for all essays.

3 PictureStories Dataset

The essays in the PictureStories dataset were collected via Write & Improve, an openly available platform for language learners. The platform is managed by CUP&A and offers automated marking and grammatical error feedback, as well as progress tracking and premium features (Yannakoudakis et al., 2018).

We are using essay answers from five picture story tasks, where language learners are tasked to *write the story that is shown in the pictures*. Each task features its own cartoon-style black and white picture story, comprising of three scenes. An example story image is shown in Figure 1 and the full set is included in Figure 6 in the Appendix. Image resolution differs slightly between images, but is around 1000 x 275 pixels.

We draw a random sample of 575 answers written by users of Write & Improve during 2024. We only select answers which are the first version submitted by the user for that story, which meet word limit requirements, and where the user has given their first language as additional metadata. An additional 145 answers were written apparently in response to one of the 5 stories, but were marked as 'off-topic' by expert human annotators engaged by CUP&A. This ensures that we have substantial representation of negative examples in our dataset.

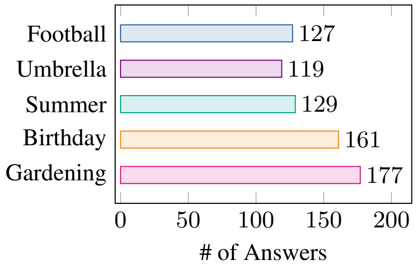


Figure 2: Number of answers for each picture story.

PictureStories contains answers from learners with over 40 different L1s.[4] Answers have an average length of 72 words, but with a high standard deviation of 54.5 due to a skewed distribution.[5] The average type-token ratio of answers for a picture story is 0.17. For story-wise statistics, see Table 6 in the Appendix. Seven answers were flagged by annotators due to profanity or because they contained personally identifiable information. These were excluded from the dataset, bringing the final total to 713 answers. For each picture story, there are between 119 and 177 answers. Exact answer counts per story are shown in Figure 2.

3.1 Marking Rubric

To cover different dimensions of task adherence, we define a marking rubric with five binary criteria. It combines aspects from the Cambridge English guidelines for teachers (Cambridge Assessment, 2020), specifically regarding the evaluation of answers given in response to picture stories, and the SAILS (King and Dickinson, 2018) and SNLI-VE (Xie et al., 2019) corpora.

Table 1 gives an overview of the marking rubric. Definitions are taken from the marking guidelines we gave to our human annotators. Rubric criteria can be grouped according to whether the picture story itself is required for evaluation: Criteria that do not depend on the specific story image evaluate the *form* of an answer. Criteria that do need the story image to be evaluated address answer *content*. Table 2 shows example answers for the *umbrella* story, which we include in Figure 4.

[4] With a slight dominance of Spanish, but also substantial presence of Portuguese, Vietnamese, Turkish, Chinese and Arabic.

[5] Answer length visualisation in Figure 7 in the Appendix.

Criterion	Description	Fleiss
Story	The answer includes linguistic features typical of stories, e.g. gives names to 'characters', uses narrative tenses, sequential adverbs, temporal phrases or 3rd person pronouns (1st person is also sometimes used).	.94
Visual	One could draw what the learner wrote without having to assume too much.	.74
Attempt	There is some relevance to the task, however tenuous.	.70
¬ Contra	Does not contradict what we see in the picture.	.61
Complete	Covers the key elements in all 3 pictures (in a reasonable order).	.63

Table 1: Overview of our marking rubric and Fleiss agreement scores between our six annotators. (¬ Contra short for non-contradictory.)

Answer Form A well-written answer can receive full marks for its *form*, even if its *content* is not at all aligned with the story image.

The task is intended to test production of a specific kind of text, namely stories. Thus, the **story** criterion assesses whether an answer is given in the form of a story. This is complemented by the determination if an answer is **visual** in the sense that it evokes a clear mental image, irrespective of the story shown in the pictures.

Answer Content *Content*-focused criteria evaluate alignment with what is shown in the story image. One criterion for which the image plays an, albeit superficial, role is whether an answer makes an **attempt** to solve the task. An answer that describes something entirely unrelated to what is shown would not fulfil this criterion. In a sense, it can be seen as a generic off-topic detection.

The first of two criteria that require a detailed comparison to the picture story is whether an answer does **not contradict** what is shown. By including this criterion, we do not desire to punish creativity. Learners are free in their interpretation of a story image and may include what could have happened before or after. Only when there is an explicit contradiction between an answer and what is shown will this lead to a negative evaluation. The final rubric element evaluates whether an answer is a **complete** rendition of the picture story. This is included as learners may stay on task, but neglect to sufficiently cover all three images of a story.

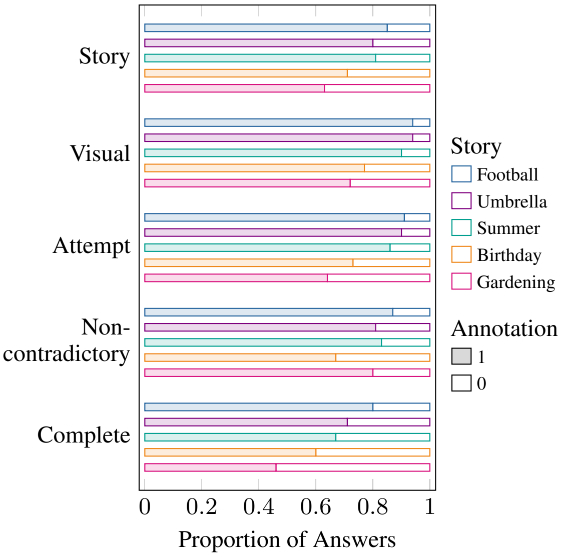


Figure 3: Annotation distribution in PictureStories.

3.2 Annotation Process

We first ran a pilot round of annotation to test the rubric and make refinements where needed. Two annotators labelled a set of 50 answers which are not part of the final dataset. Once agreement was high (>.9 QWK for all variables) we proceeded with annotation of the answers in PictureStories.

Six annotators each labelled the full dataset. Table 1 shows their overall agreement. For detailed, pairwise agreement see Table 5 in the Appendix. Roughly, there are three tiers: Agreement for *story* is highest (.94 Fleiss), followed by *visual* and *attempt* (.74 and .70) and then *non-contradictory* and *complete* (.61 and .63, respectively). For the *non-contradictory* annotation, one annotator (r2 in Table 5) tended to be very lenient, which brings down agreement for this criterion.

3.3 Label Distribution

We derive the ground truth label from the six annotations via majority vote. In case of ties, we assign the lower rating, i.e. an annotation of 0. Figure 3 shows the story-wise label distribution in the final dataset. Overall, there is a tendency for criteria to be met.

4 Experimental Setup

To test how well our rubric criteria can be predicted automatically, we compare prompting three openly available LLMs to finetuning different model types.

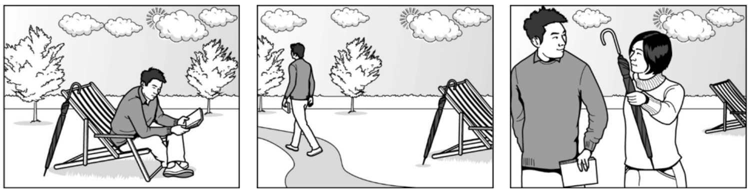


Figure 4: Story image in PictureStories for which example answers are shown below.

Variable	Positive Example	Negative Example
Story	One time Carlos was reading his favorite book in the park, [...]	In these pictures we can see a man sits in a chair while he's reading [...]
Visual	[...] Sarah, who was near him, saw the umbrella and decided to carry it to him.	[...] that day was so hard with the clim but its okay when I am reading [...]
Attempt	[...] The Sun getting hoter so he move, and he left his umrella [...]	[...] my faithful friend Bug who help me survive in this new world of giant lizards [...]
Non-contradictory	John went to park and spent a couple of hours reading a book. It was a great book, [...]	Rob was reading a book in the park, but the rain started. [...]
Complete	There are three pictures, from left to right. A man is setting down on a chair and reading a book and an umbrella is behind him. In the second picture, he is leaving his place and forget his umbrella. In the third picture, a girl brign his umbrella and gives it to him.	In the first pisture, I see a man readin a book while sunbathing in the garden. He looks very confused. In the second picture, the man is walking garden. In the last picture, the man and his friend meet and walk while chatting.

Table 2: Example learner answers for each criterion of our marking rubric, written in response to the *umbrella* story shown above.

4.1 Models

Finetuning without images As a reference point, we fit a **logistic regression (LR)** classifier (scikit-learn version 1.6.1). As features we use a *CountVectorizer* with lowercased unigrams and bigrams. Other than setting *max_iter* to 1000, all parameters are left at their respective default values. To complement the shallow LR classifier, we run experiments with two deep learning models, **BERT** (Devlin et al., 2019) and **SBERT** (Reimers and Gurevych, 2019). We finetune BERT (*bert-base-uncased*[6], transformers version 4.49.0) with a classification head for 30 epochs using a batch size of 16. For SBERT (Reimers and Gurevych, 2019) (*all-MiniLM-L6-v2*, sentence-transformers version 4.0.1), we follow the setup described in Bexte et al. (2022). They propose a finetuning of SBERT using answer pairs built from the pool of training data. Thereby, the model learns to predict the similarity in the scores of two answers. At inference, a kNN-like search in the embedding space is used to compare test answers to training answers. The score for which training answers are most similar to the test answer is predicted. We also report performance of the pretrained model without any finetuning, i.e. performing inference directly with the pretrained SBERT model.

Finetuning with images All previously described models are trained exclusively on student answers, omitting the story image. To test finetuning that includes images, we use **ALIGN**[7] (*kakaobrain/align-base*). ALIGN embeds image and text into a shared space. Analogous to finetuning BERT with a classification head, we do the same with the textual ALIGN embeddings. This establishes the text-only performance of ALIGN. To incorporate the images, we test two options: **concatenation** and **subtraction**. In the former, we concatenate the embedding of the student answer and the story image.[8] This is then fed through a classification head. In subtraction, we embed story image and answer and then subtract the visual embedding from the textual one. With this, we attempt to capture the essence of an answer, independent of the context given by the image. The difference embedding is then again passed to a classification head. In both conditions, we train for 30 epochs with a batch size of 16 and a learning rate of 1e-05.

[6] We also tested *answerdotai/ModernBERT-base*, but found this model to performe worse.

[7] We also tested CLIP, but found ALIGN to perform better.

[8] We also tested embedding the story as three separate images, but found this to perform slightly worse.

LLMs We test three vision-and-text LLMs that are openly available via Hugging Face (Wolf et al., 2020). In a preliminary study, we tested how well models can grasp the content of the picture stories in PictureStories, and how strong their concept of a *story* is. To this end, we prompted five different vision-and-language LLMs (LLaVA, Qwen, Idefics, BLIP, Pixtral) to take the role of the language learner and *write the story that is shown in the pictures*, just like language learners are asked to. This revealed Qwen and LLaVA to be capable of producing extensive fairytale-like texts. Another model that stood out was Idefics, as it produced brief, descriptive answers that accurately represented what is shown in the story images.

Thus, our analysis focuses on LLaVA (*llava-hf/llava-v1.6-vicuna-7b-hf*), Qwen (*Qwen/Qwen2.5-VL-3B-Instruct*) and Idefics (*HuggingFaceM4/Idefics3-8B-Llama3*), all of which are prompted in the default configuration (transformers version 4.46.0).

Prompt Design For each learner answer and criterion, an individual request is put towards an LLM. The description of the respective criterion is taken from our annotation guidelines. All runs are done twice, once inputting both the learner answer and the story image and once just the learner answer. This *blind* condition enables us to assess the effect of adding example answers to the prompt vs. including the story image.

With LLMs, even minor changes in the prompt can affect results (Sclar et al., 2023). Thus, we use a modular prompt design that introduces slight variations, each of which should not influence the model's predictions. These modifications are whether we speak of *images* or *pictures*, *categories* or *criteria* that are *satisfied* or *fulfilled* and a slight variation on the statement immediately before presenting the learner answer the model is asked to make a prediction on. The full prompt design is included in Figure 9 in the Appendix. In our results, we report average performance across prompt variations. Performance ranges are shown in Figure 10 in the Appendix.

4.2 Data Split

From a teacher's perspective, it is desirable to not have to manually annotate large quantities of data whenever a new task (in our case a new picture story) is introduced. Therefore, our experiments focus on settings where limited training data is available for the test story.

For the models we finetune (LR, BERT, SBERT), we run experiments in a *cross-story* setup. To split the data, answers for one picture story are taken as test data. For each of the other stories, a random sample of ten percent of the answers is set aside for validation. The remaining answers are used as training data. Thus, models will have to abstract from the four stories present during training to a previously unseen fifth story at test time.

As a reference point to this cross-story setting, we also report *within-story* performance, i.e. fitting a dedicated model based on the answers for a single picture story. This scenario allows for the model to adapt to the content relevant to an individual story, e.g. that the man forgets his umbrella in the story depicted in Figure 4. Thus, the image contents will in large part be covered by the training answers themselves, which is why we do not include the image in fitting the model. Instead, we run a simple leave-one-out cross validation with the logistic regression model.

We always train dedicated models for each of our five target labels, just like we always prompt LLMs to predict just one of our criteria. This tests model ability to adapt to one specific target concept.

With the LLMs we test, we explore zero-shot prompting and few-shot prompting. We define one shot as a pair of one positive and one negative example, i.e. one example answer that meets the respective criterion and one that does not. Thus, when we speak of *n*-shot prompting we include *n* positive and *n* negative examples in the prompt. As reference examples we prioritise those that all annotators agreed upon. We make sure to pair each learner answer with the same examples in prompting the different models, so as not to unfairly disadvantage a model by an unlucky draw of reference answers. Regarding the distinction of within-story or cross-story prediction, the few-shot prompting is *within-story* in the sense that the examples that are included in the prompt belong to the same story as the answer that should be classified.

Evaluation To account for the label imbalance in the data, we use macro-averaged F1 to measure performance. We always report average performance across all five stories. For finetuned models, performance is averaged across five runs. When an LLM returns an answer that does not contain *True* or *False*, i.e. fails to classify an answer[9], we remap

[9] See Table 7 in the Appendix for an overview of how often this is the case.

this as a prediction of the respective *incorrect* class. Therefore, it will be interpreted as predicting 1 when the gold label is 0, and vice versa.[10]

Infrastructure Experiments ran on an NVIDIA A100 for a total of around 100 GPU hours.

5 Results

Table 3 gives an overview of our results. Within-story performance is at a solid level of over .8 F1 for all five criteria. This demonstrates the beneficial effect of having training answers for the same picture story a model is evaluated on, especially considering that it uses about 75% less training data than cross-story scoring.

Finetuned models Models finetuned cross-story outperform the within-story model for *story* and *visual*. This is in line with our earlier discussion of the picture story-independent nature of these form-focused criteria. Results show that it is beneficial to have more cross-story training data over a smaller set of within-story answers when evaluating form.

Dependence on the contents of a story image for the other three criteria is mirrored by a performance drop of cross-story models over within-story ones. Especially for *complete*, performance drops substantially, from .82 within-story to under .6 cross-story, underlining the advantage of the within-story model's adaptation to the content of a specific story.

Interestingly, comparing the performance of SBERT with and without finetuning, performance for *story* and *visual* increases after finetuning, but decreases for the content-focused criteria. Thus, adapting the model to answers for the other stories skews it towards them to an extent that harms performance on answers for the held-out test story.

For the ALIGN model, there is no effect of including the story image. This emphasizes the difficulty of our story depictions. They have a relatively low resolution and do not match the predominantly photographic training data of ALIGN.

LLMs Other than ALIGN, LLMs benefit from the additional context given by the story image. With the exception of LLaVA for *story* and *visual*, there is a consistent performance increase when the image is included in zero-shot prompting. This effect is less pronounced in 3-shot prompting, as some of the information included in the image will be covered by the example answers included in the prompt.

For the two form-focused criteria *story* and *visual*, LLMs fall short of the cross-story finetuned models. Their strength lies in predicting the content-focused variables. For these, Qwen comes close to within-story training. Still, this is impressive considering that just three examples per outcome are included in the prompt.

5.1 Prompting with Image vs. Examples

To grasp the effects of including the story image and/or an increasing number of example answers in the prompt, we visualise this relation in Figure 5. Noting the relatively good performance of the SBERT model without any finetuning in our cross-story scoring, we also include performance of this model when being fed the same examples.

Comparing the much simpler SBERT model to the LLMs shows it to outperform LLaVA in almost all cases. Thus, it should not be underestimated how informative the few-shot examples included in a prompt are, even to a much less complex model.

Especially for Qwen, there is an impressive benefit of providing examples. In zero-shot prompting, including the story image has a substantial positive effect on performance. With the addition of just one example per class, performance of prompting without the image catches up. The only exception is *attempt*, but even here there is a pronounced benefit of the added examples. Beyond one answer per outcome, adding further examples has much less of an incremental effect on performance.

5.2 Relative Difficulty of Stories

Seeing as previous results reported average performance across all five picture stories, we now take a look at the individual stories. A t-SNE visualisation of answers embedded with the SBERT model shows a relatively uniform separation into stories.[11]

As Qwen had shown the overall most solid performance, we break down per-story performance for this model. Table 4 compares the zero-shot setting with inclusion of the story image to 1-shot prompting without including the image.

For the zero-shot prompting condition, there is a relatively clear trend of which story performs worst/best. For all rubric criteria, it is the *umbrella* story that does worst. For all but the *non-*

[10] The alternative would be to introduce a third label that is not present in the gold standard. However, this will lead to the F1 value for this class to be 0, but still make up a third of the macro-averaged F1, even if occurs just once.

[11] See Figure 8 in the Appendix for this visualisation.

Model	w/ Image	Story	Visual	Attempt	¬ Contra	Complete	Avg.
				Performance References			
Majority Baseline	✗	.43	.46	.44	.42	.37	.42
Within-Story (LR)	✗	.83	.83	.91	.87	.82	.85
				Cross-Story			
LR	✗	.83	.86	.71	.61	.39	.68
BERT	✗	.91	.92	.65	.69	.53	.74
SBERT (no finetuning)	✗	.71	.83	.74	.69	.57	.71
SBERT	✗	.89	.92	.69	.64	.50	.73
ALIGN (Text-only)	✗	.88	.89	.78	.70	.58	.76
ALIGN (Concat)	✓	.88	.89	.75	.70	.59	.76
ALIGN (Subtract)	✓	.88	.87	.77	.71	.59	.77
				Zero-Shot			
Idefics	✗	.57	.51	.62	.63	.63	.59
Idefics	✓	.62	.68	.82	.78	.71	.72
LLaVA	✗	.74	.67	.49	.60	.55	.61
LLaVA	✓	.63	.62	.71	.66	.55	.64
Qwen	✗	.52	.61	.64	.46	.46	.54
Qwen	✓	.67	.78	.89	.78	.70	.77
				3-Shot			
Idefics	✗	.74	.67	.67	.78	.69	.74
Idefics	✓	.74	.71	.80	.72	.73	.74
LLaVA	✗	.59	.56	.50	.54	.43	.52
LLaVA	✓	.52	.69	.52	.68	.60	.60
Qwen	✗	.79	.87	.84	.84	.74	.82
Qwen	✓	.76	.82	.87	.82	.74	.80

Table 3: Macro-averaged F1 performance, grouped by the kind of data predictions are based on. Depending on the model, scoring is done purely on answers (✗) or also incorporates the image (✓) (¬ Contra short for non-contradictory). Green-blue shading indicates lower-to-higher F1 values.

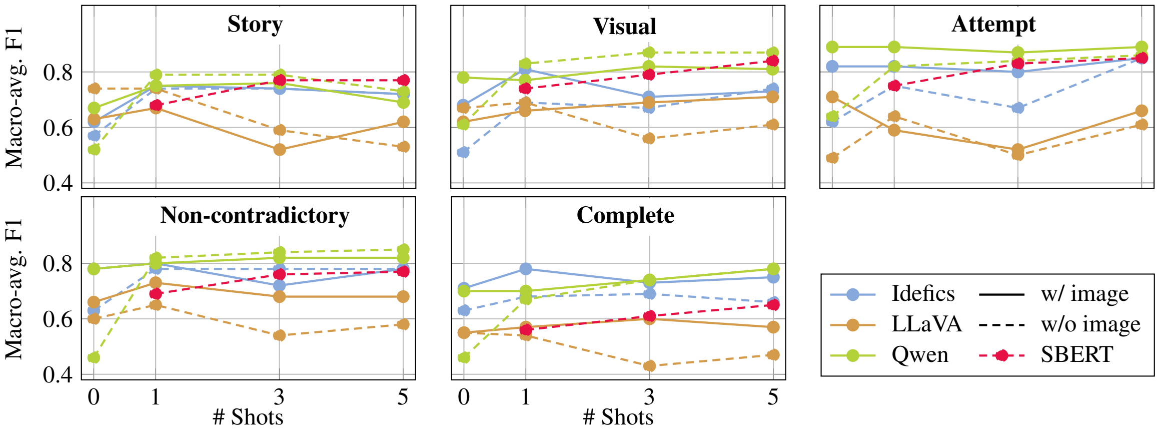


Figure 5: Model performance depending on how many examples are incorporated in the prompt (x axis). Dashed lines indicate that story images were not shown.

contradictory criterion, it is the *birthday* story that does best. There is much more variation in relative story performance for the 1-shot prompting condition. Depending on the criterion, all but the *gardening* story are at least once the lowest-performing one. Similarly, all but the *football* story are at least once the best-performing one.

Variation in model performance on the 5 different picture stories is explained to some extent by the quality of examples selected for few-shot learning. For zero-shot prompting, the model might have an easier time processing some over other story images, which would be in line with the more consistent pattern of which story exhibits the best/worst performance.

6 Discussion

We publish PictureStories, a dataset of L2 English learner answers written about image-based narratives under a custom license that permits non-commercial use. We develop a marking rubric that

Picture Story	# shots	w/ image	**Story**	**Visual**	**Attempt**	**¬ Contra**	**Complete**
Birthday	0	✓	.74	.79	.92	.84	.72
	1	✗	.79	.76	.86	.88	.73
Football	0	✓	.64	.77	.87	.74	.65
	1	✗	.77	.82	.84	.74	.36
Gardening	0	✓	.61	.75	.92	.85	.71
	1	✗	.83	.85	.84	.86	.75
Summer	0	✓	.65	.76	.87	.70	.72
	1	✗	.73	.85	.70	.80	.66
Umbrella	0	✓	.57	.71	.79	.62	.61
	1	✗	.71	.81	.73	.67	.77

Table 4: Macro-averaged F1 performance of zero-shot prompting Qwen with inclusion of the story image vs. 3-shot prompting the model without including the image, broken down into stories. (¬ Contra short for non-contradictory.)

evaluates the task adherence of learner answers, covering both form and content. Six annotators rate the entire dataset.

Aiming for a method of automated scoring that does not rely on extensive amounts of manually labelled training data for a target story, we compare different scoring models. Results show cross-story finetuned models to be the best-performing method to evaluate answer form. For content-focused criteria, prompting Qwen emerges as a robust method that comes close to within-story training. For the other LLMs we test (LLaVA and Idefics), few-shot performance is often matched by running a simple inference with a pretrained SBERT model.

When using Qwen to evaluate answers, including just one positive and one negative example in the prompt is in many cases just as helpful as passing along the story image instead. The condition in which Qwen gives the overall best performance is even to use three positive and three negative examples without inclusion of the story image.

Overall, our results are a reminder that the suitability of a scoring model depends on the target variable. For LLMs, it is worthwhile to consider the trade-off between the inclusion of an image vs. example answers; and even if few-shot performance can look promising, a simpler model might do just as well with the same examples.

Limitations

Even though we do include minor variations in the prompts we use to score answers with LLMs, there are many more ways of phrasing this request. As we are already seeing in our small prompt variations, this will influence results.

The vast majority of data that visual models are pretrained on are photographs (Jia et al., 2021; McKinzie et al., 2025). The images in PictureStories are black and white cartoon-style illustrations of a story in three scenes. This makes them unusual in the sense that they are a) not photographs and b) technically three individual images. Thus, it may be worthwhile to explore preprocessing of the images into a style that is more familiar to models. Additionally, one could explicitly point the model towards the fact that there are three separate scenes or even present them one by one.

Ethical Considerations

While images are exciting due to their language-invariance, their interpretation is not free from cultural influence (Masuda and Nisbett, 2001). Not anticipating this might lead to unexpected answers. Similarly, learners may come up with creative interpretations of the visually presented story. Overall, it is advisable to err on the side of leniency when evaluating content, as not to unnecessarily disadvantage such answers. Since the overall aim of the task is to foster language learning, creative answers should not be punished, but rather encouraged. If students are overly concerned with meeting the expectations of the examiners, this may lead to the development of test taking strategies to *play it safe* (Xu and Wu, 2012), which goes against the aim of cueing free language production.

Annotations were carried out by 6 expert annotators who are engaged by the dataset providers. The annotation work was carried out as part of their contractual roles, and they receive fair remuneration for their work from the dataset providers in line with UK law and market rates.

In labelling the answers in PictureStories, the six annotators screened for personally identifiable information and profanity. Seven answers were subsequently excluded. Since none of the remaining answers was flagged by any of our annotators, we are confident that they do not contain ethically problematic content.

Under the Terms of Use of Write & Improve, users license the use of their submitted texts for research purposes.

Acknowledgments

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany. The ALTA Institute is supported by Cambridge University Press & Assessment. We thank colleagues at CUP&A, in ALTA and the Computer Laboratory for their support and feedback.

References

Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Graded Relevance Scoring of Written Essays with Dense Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1329–1338, New York, NY, USA. Association for Computing Machinery.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Timo Baumann, Korbinian Eller, and Natalia Gagarina. 2024. BERT-based Annotation of Oral Texts Elicited via Multilingual Assessment Instrument for Narratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 99–104, Miami, Florida, USA. Association for Computational Linguistics.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2024. Assessing how accurately Large Language Models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence*, page 100353.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring - how to make S-BERT keep up with BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123, Seattle, Washington. Association for Computational Linguistics.

Frank Boers. 2018. Picture prompts and some of their uses. *Language Teaching Research*, 22(4):375–378. Publisher: SAGE Publications.

Cambridge Assessment. 2020. *A2 Key for Schools: Handbook for Teachers for Exams*. Cambridge English Qualifications. Cambridge University Press.

Ronan Cummins, Helen Yannakoudakis, and Ted Briscoe. 2016. Unsupervised modeling of topical relevance in L2 learner text. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 95–104, San Diego, CA. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

D. Higgins, J. Burstein, and Y. Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

Pengcheng Huang, Li Li, Chunyan Wu, Xiaoqian Zhang, and Zhigui Liu. 2023. A Study of Sentence-BERT Based Essay Off-topic Detection. In *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, CNIOT '23, pages 515–519, New York, NY, USA. Association for Computing Machinery.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR. ISSN: 2640-3498.

Levi King and Markus Dickinson. 2016. Shallow Semantic Reasoning from an Incomplete Gold Standard for Learner Language. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 112–121, San Diego, CA. Association for Computational Linguistics.

Levi King and Markus Dickinson. 2018. Annotating picture description task responses for content analysis. In *Proceedings of the Thirteenth Workshop on*

Innovative Use of NLP for Building Educational Applications, pages 101–109, New Orleans, Louisiana. Association for Computational Linguistics.

Katsunori Kotani, Takehiko Yoshimi, Hiroaki Nanjo, and Hitoshi Isahara. 2011. Compiling learner corpus data of linguistic output and language processing in speaking, listening, writing, and reading. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1418–1422, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Christine Köhn and Arne Köhn. 2018. An Annotated Corpus of Picture Stories Retold by Language Learners. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 121–132, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Simon Masloch, Doreen Scholz, Eva Belke, and Stefanie Dipper. 2019. The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*, 51(4):1889–1918.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems*, volume 36, pages 71683–71702. Curran Associates, Inc.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Takahiko Masuda and Richard E. Nisbett. 2001. Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81(5):922–934. Publisher: American Psychological Association.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2025. MM1: Methods, Analysis and Insights from Multimodal LLM Pre-training. In *Computer Vision – ECCV 2024*, pages 304–323, Cham. Springer Nature Switzerland.

OpenAI. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Isaac Persing and Vincent Ng. 2014. Modeling Prompt Adherence in Student Essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Marek Rei. 2017. Detecting off-topic responses to visual prompts. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yashad Samant, Lee Becker, Scott Hellman, Bradley Behan, Sarah Hughes, and Joshua Southerland. 2025. Automatic Detection of Inauthentic Templated Responses in English Language Assessments. ArXiv:2509.08355 [cs].

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Swapna Somasundaran and Martin Chodorow. 2014. Automated Measures of Specific Vocabulary Knowledge from Constructed Responses ('Use These Words to Write a Sentence Based on this Picture'). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Baltimore, Maryland. Association for Computational Linguistics.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated

Scoring of Picture-based Story Narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, Denver, Colorado. Association for Computational Linguistics.

Kento Tanaka, Taichi Nishimura, Hiroaki Nanjo, Keisuke Shirai, Hirotaka Kameko, and Masatake Dantsuji. 2022. Image Description Dataset for Language Learners. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6814–6821, Marseille, France. European Language Resources Association.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hanna Woloszyn and Benjamin Gagl. 2025. Can Large Language Models (LLMs) Describe Pictures Like Children? A Comparative Corpus Study. *arXiv preprint*. ArXiv:2508.13769 [cs].

Jane Wottawa and Martine Adda-Decker. 2016. French learners audio corpus of German speech (FLACGS). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3215–3219, Portorož, Slovenia. European Language Resources Association (ELRA).

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment Task for Visually-Grounded Language Learning. *arXiv:1811.10582 [cs]*.

Yun Xu and Zunmin Wu. 2012. Test-taking strategies for a high-stakes writing test: An exploratory study of 12 Chinese EFL learners. *Assessing Writing*, 17(3):174–190.

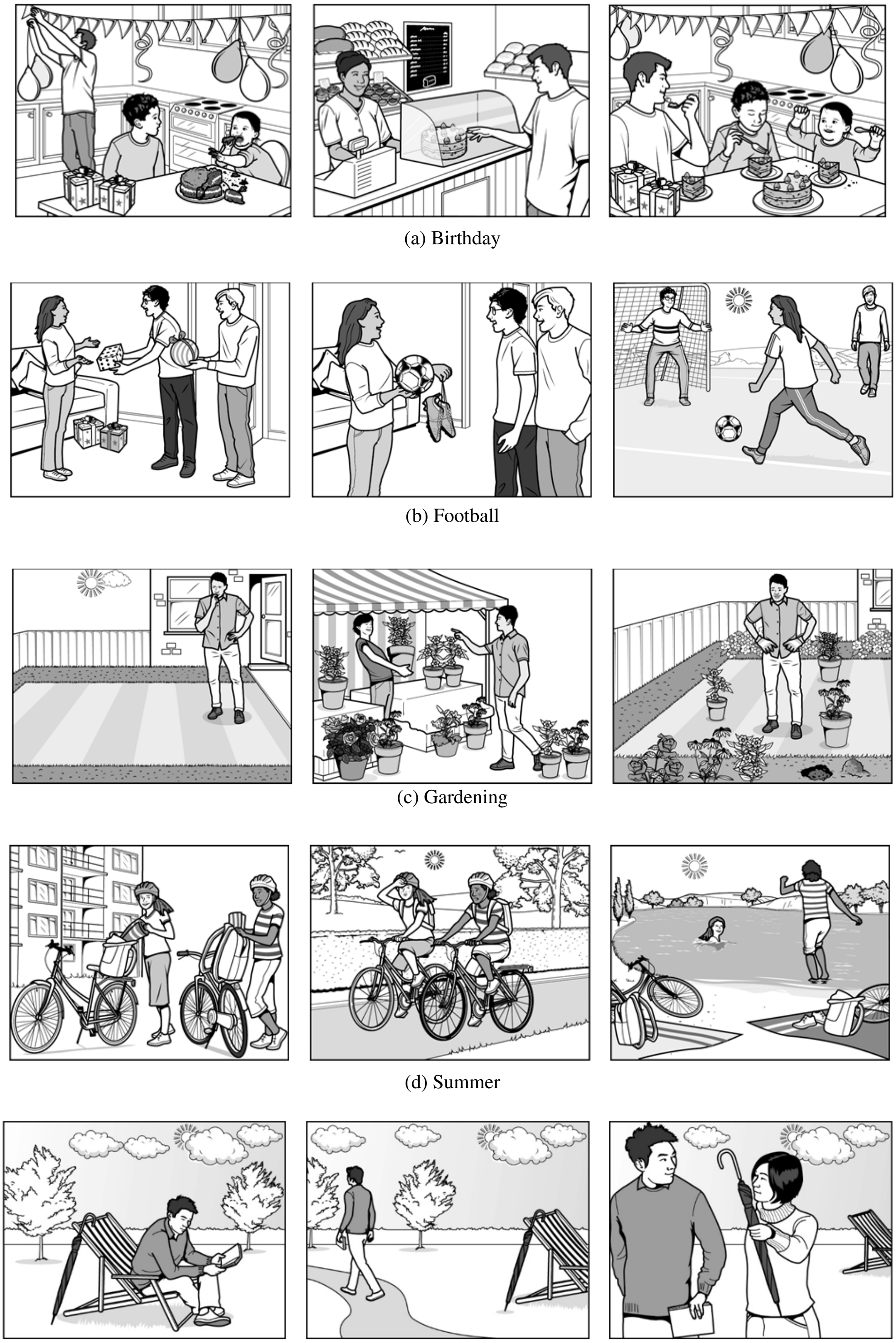
Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

A Supplementary Material

In this Appendix, we include material and additional results that are intended to increase transparency.

With respect to the dataset, Figure 6 shows the full set of story images in PictureStories. Table 5 gives a detailed breakdown of annotator agreement. In Table 6, we show the story-wise label distribution, average answer length and type-token ratio. This is complemented by a depiction of answer lengths in Figure 7. Figure 8 gives a t-SNE overview of answers embedded with the pretrained SBERT model.

Regarding our experiments, Figure 9 has a pseudo code representation of the prompt design we use for score prediction with LLMs. Figure 10 visualises performance variation across the different prompt versions. Table 7 lists how often the individual LLMs failed to produce a response of *True* or *False*.



(a) Birthday

(b) Football

(c) Gardening

(d) Summer

(e) Umbrella

Figure 6: Overview of the story images in PictureStories.

	r2	r3	r4	r5	r6	vote
r1	.94	.96	.93	.95	.97	.99
r2	–	.94	.92	.93	.92	.95
r3		–	.92	.94	.95	.97
r4			–	.93	.93	.94
r5				–	.93	.96

(a) Attempt

	r2	r3	r4	r5	r6	vote
r1	.79	.60	.85	.74	.81	.87
r2	–	.62	.81	.76	.82	.87
r3		–	.63	.67	.64	.69
r4			–	.81	.86	.89
r5				–	.82	.86

(b) Story

	r2	r3	r4	r5	r6	vote
r1	.86	.75	.61	.82	.83	.92
r2	–	.70	.57	.83	.79	.89
r3		–	.54	.72	.72	.79
r4			–	.62	.58	.66
r5				–	.79	.89

(c) Visual

	r2	r3	r4	r5	r6	vote
r1	.37	.81	.84	.68	.88	.95
r2	–	.28	.27	.26	.32	.40
r3		–	.75	.65	.83	.85
r4			–	.59	.84	.84
r5				–	.65	.69

(d) Non-contradictory

	r2	r3	r4	r5	r6	vote
r1	.81	.56	.66	.61	.78	.82
r2	–	.58	.63	.57	.77	.81
r3		–	.54	.60	.56	.70
r4			–	.60	.73	.76
r5				–	.60	.72

(e) Complete

Table 5: Pairwise QWK agreement of the six annotators for the different criteria. The rightmost column (**vote**) shows agreement with the adjudicated annotations. Values are highlighted according to the strength of agreement classification of Landis and Koch (1977) as fair , moderate , substantial or almost perfect agreement.

		# Meeting Criterion						
Picture Story	**Total #**	Story	Visual	Attempt	Non-contr.	Complete	**Avg. Length (SD)**	**Type-Token Ratio**
Birthday	161	114	108	117	108	97	82.0 (72.8)	.18
Football	127	108	119	115	110	102	69.9 (65.0)	.16
Gardening	177	112	128	114	105	81	67.2 (48.6)	.18
Summer	129	104	116	111	107	86	65.6 (38.8)	.18
Umbrella	119	95	112	107	96	84	65.0 (29.0)	.15

Table 6: Answer counts, label distribution, average answer length in tokens and type-token ratio, broken down into the individual stories.

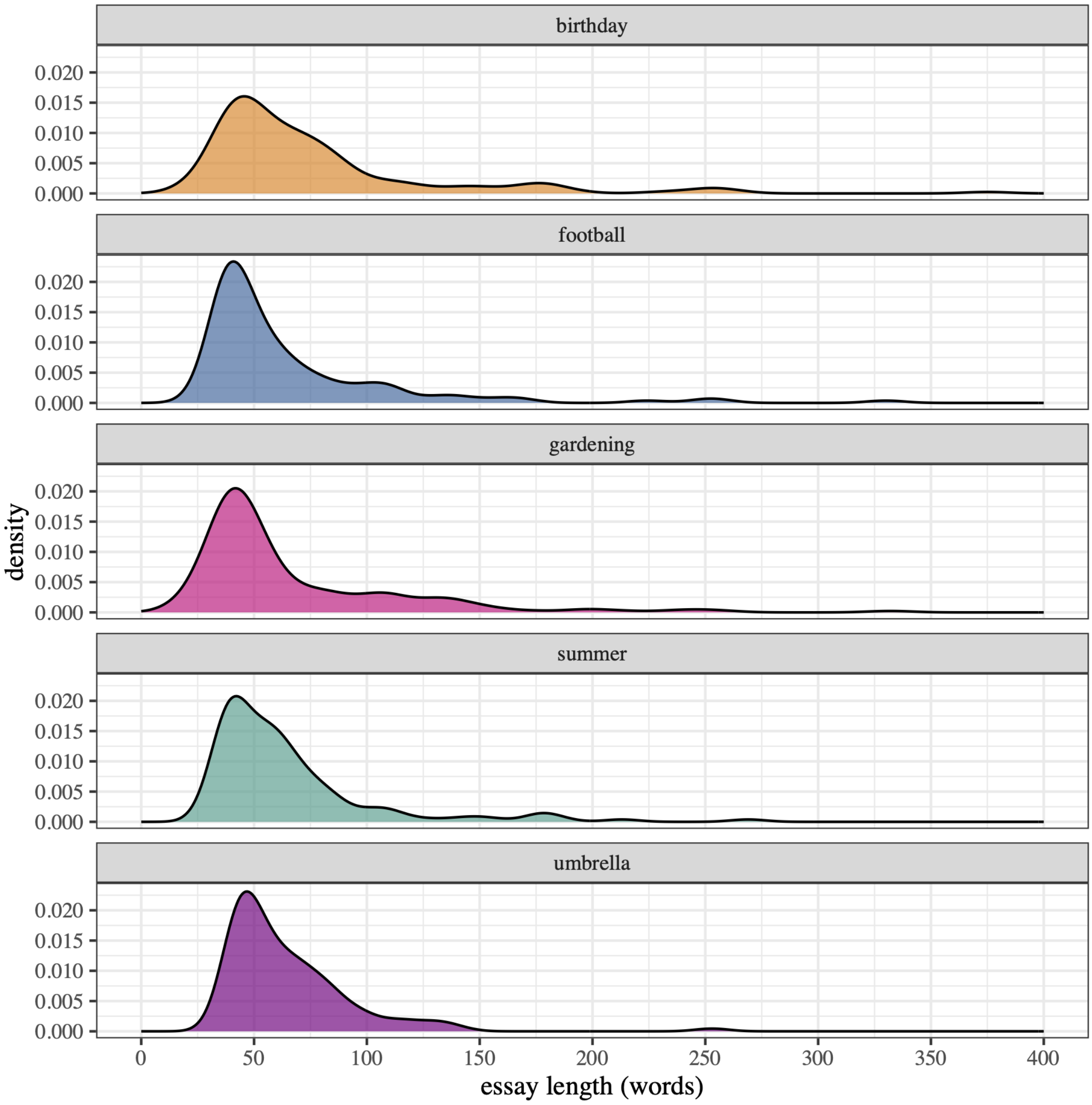


Figure 7: Density plot of answer lengths for each of the 5 picture stories in PictureStories. The longest answer in the dataset is 589 words, but we clipped the plot at 400 words since there is relatively low density beyond this point.

Model	0-shot	1-shot	3-shot	5-shot
Idefics	99	53	44	47
LLaVA	0	0	30	1473
Qwen	0	0	0	0

Table 7: Number of cases where the respective LLM is not conforming to the requested answer of *True* or *False*. This is out of a total of 114080 predictions in each run.

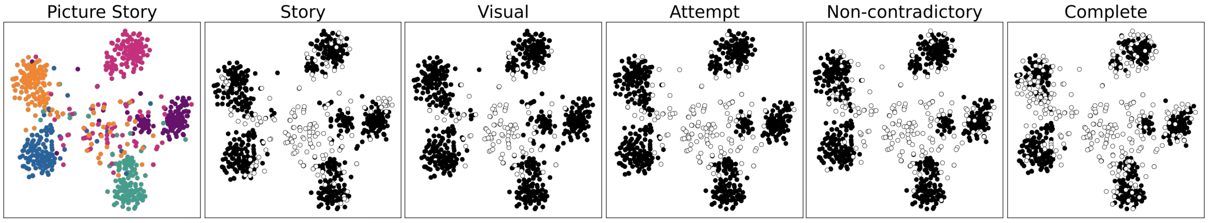


Figure 8: t-SNE visualisation of the dataset, differentiating the stories in PictureStories: birthday, football, gardening, summer, and umbrella. Each answer was encoded with the pretrained SBERT model. Black circles signify fulfilment of the respective criterion. There is a clear separation of answers into the five stories. Negative examples for the different criteria (white circles) tend to skew towards the middle. *Complete* had proven to be the most challenging of the five variables in our experiments. As can be gathered from Figure 3, the proportion of negative examples is highest for this criterion, and as can be seen here, negative examples cluster closely to positive examples for the same image.

```
for BLIND in [True, False]:

  for ELEMENTS in ["images", "pictures"]:

    for VARIABLE_REF in ["criterion", "condition"]:

      for VARIABLE_OK in ["satisfy", "fulfil"]:

        for ANSWER in ["This is the learner answer you have to evaluate:\n LEARNER_ANSWER",
                     "This is the what the learner wrote and must be evaluated:\n LEARNER_ANSWER"]:

          IF BLIND:
            "A language learner was asked to write the story that is shown in a set of ELEMENTS."

          ELSE:
            "A language learner was asked to write the story that is shown in these ELEMENTS."

          "You are a teacher and have to decide whether their answer VARIABLE_OKs the following
          VARIABLE_REF:"

          # Answerhood
          "It is an attempt to do the task."
          "This VARAIABLE_REF is fulfilled if there is some relevance to the task, however tenuous."

          OR # Story
          "It is a story (however clumsy) or 'storylike'."
          "This VARIABLE_REF is VARIABLE_OKd if the answer includes linguistic features typical of
          stories, e.g. gives names to 'characters', uses narrative tenses, sequential adverbs,
          temporal phrases or 3rd person pronouns (1st person is also sometimes used)."

          OR # Visual
          "It evokes a clear mental image."
          "This VARIABLE_REF is VARIABLE_OKd if one could draw what they wrote without having to assume
          too much."

          OR # Non-contradictory.
          "It does not contradict what we see in the ELEMENTS."

          OR # Complete
          "It covers they key elements in all 3 ELEMENTS (in a reasonable order)."

          "Please answer with either 'True' or 'False'."

          IF FEW_SHOT:
            "Below are some exemplary answers and whether they VARIABLE_OK the VARIABLE_REF:"

            for POS in POS_EXAMPLES:
              "'True': POS"

            for NEG in NEG_EXAMPLES:
              "'False': NEG"

          ANSWER
```

Figure 9: Prompting strategy used for LLM-based evaluation of answers.

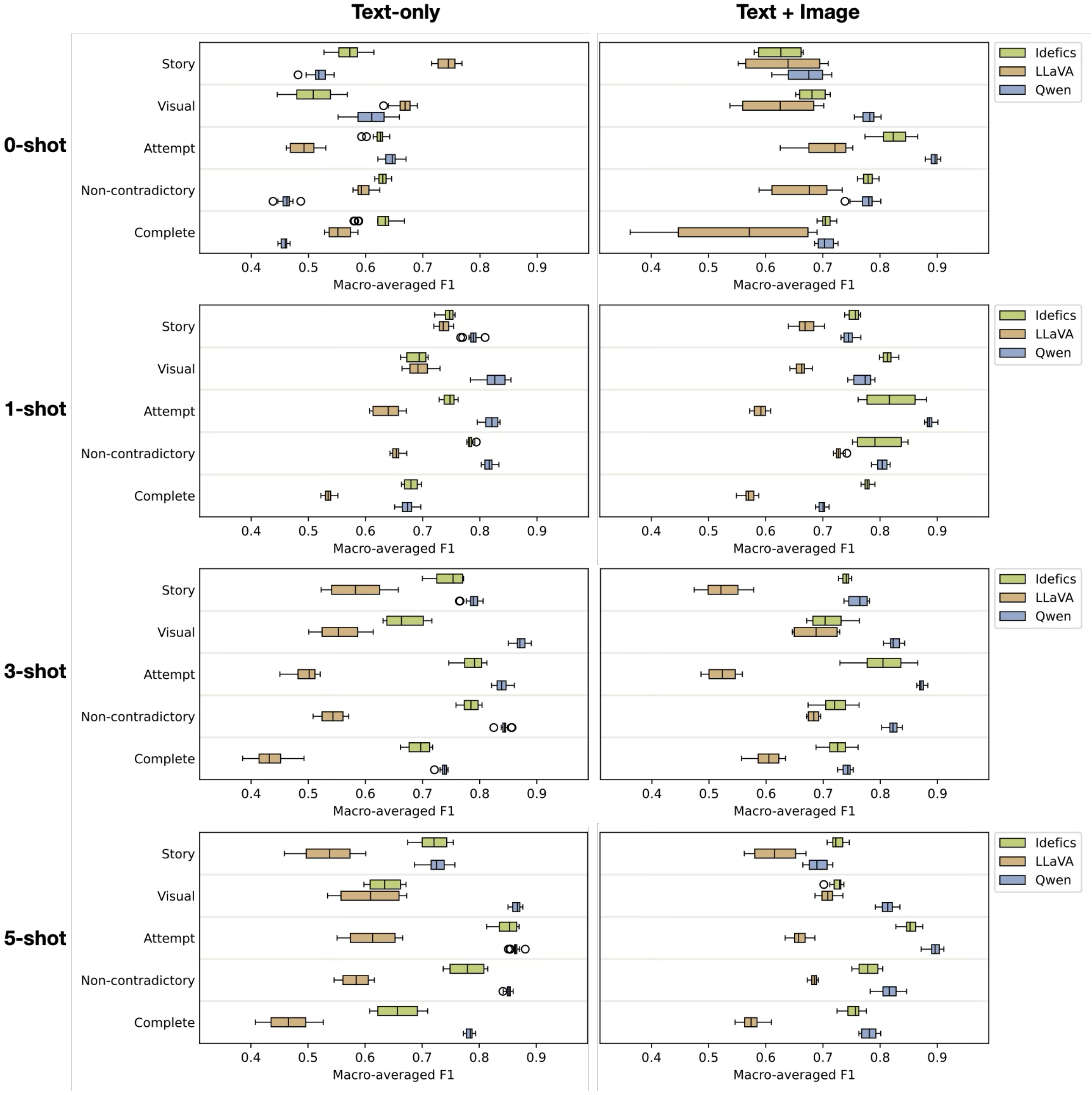


Figure 10: Prediction performance of the LLMs when adding *n* positive and *n* negative examples to the prompt (rows). Right column shows performance when the story image is included in the prompt, left column when it is not. Each boxplot shows the performance range across our 16 prompt variations. In the main paper, we report average performance.