

# From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation

**Ali Malik**  
Stanford University  
Stanford, CA

**Stephen Mayhew**  
Duolingo  
Pittsburgh, PA

**Chris Piech**  
Stanford University  
Stanford, CA

**Klinton Bicknell**  
Duolingo  
Pittsburgh, PA

## Abstract

We study the problem of controlling the difficulty level of text generated by Large Language Models (LLMs) for contexts where end-users are not fully proficient, such as language learners. Using a novel framework, we evaluate the effectiveness of several key approaches for this task, including few-shot prompting, supervised finetuning, and reinforcement learning (RL), utilising both GPT-4 and open source alternatives like LLama2-7B and Mistral-7B.

Our findings reveal a large performance gap between GPT-4 and the open source models when using prompt-based strategies. However, we show how to bridge this gap with a careful combination of finetuning and RL alignment. Our best model, CALM (CEFR-Aligned Language Model), surpasses the performance of GPT-4 and other strategies, at only a fraction of the cost. We further validate the quality of our results through a small-scale human study.

## 1 Introduction

Large Language Models (LLMs) are powerful tools for content generation. However, these models often output text at a native level of speech (Figure 1, top). This makes LLMs challenging to use for contexts where the end users are not fully proficient, such as for language learners, young children, or non-native speakers. When generating content for these use cases, we need the ability to control the proficiency level of the generated text.

In this work, we formally define the Proficiency Control Task (PCT): a new framework that assesses a model’s ability to modulate language proficiency level, while also generating high-quality content consistent with given instructions. We evaluate models with respect to the three essential criteria: (1) *ControlError* which measures how close the generated text is to the target proficiency, (2) *QualityScore*, which measures relevance of the text to the instructions, and (3) *Cost* which measures the resource-intensiveness of the approach.

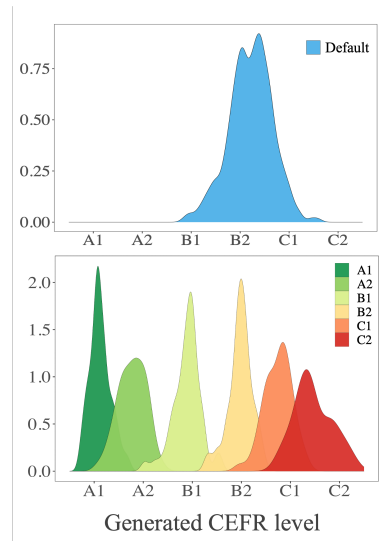


Figure 1: **(top)** GPT-4 generates content at a native proficiency level. **(bottom)** Results from our CALM proficiency control model for different target levels.

Using this evaluation framework and the TinyStories dataset (Eldan and Li, 2023), we investigate several key approaches to the PCT on the task of short story generation from a plot summary.

**Prompt-based approaches** First, we thoroughly explore the space of few-shot, prompt based strategies with OpenAI’s GPT-4 and open source alternatives (Section 6). Our findings demonstrate the strong capability of GPT-4 at the PCT, resulting in low *ControlError* and high *QualityScore*. We also identify an improvement in *ControlError* as prompts are made more complex, resulting in better proficiency control at the cost of more tokens.

Although GPT-4 is successful at the PCT, it is a proprietary model and its generations are several times more costly than open source alternatives. However, we find instruction-tuned models like LLama-2-7b and Mistral-7b perform poorly at the PCT through prompting.

**Finetuning open source models** To bridge the gap between open source models and GPT-4, we turn to supervised finetuning approaches from the controllable text generation literature (Keskar et al., 2019; Stowe et al., 2022). Specifically, we use the outputs of an effective GPT4 prompting strategy to generate data for the PCT that can be used to directly train open source models.

Using this data, we are able to finetune LLaMa2-7b and Mistral-7b to come significantly closer in performance to GPT-4 at the PCT (Section 7). Moreover, we show how additional training with Proximal Policy Optimisation (PPO) can further align the outputs of these models with the desired proficiency levels. Our best such model, we call CALM (CEFR-Aligned Language Model), has a *ControlError* equal to that of GPT-4, at only a fraction of the cost.

**Boosting PCT models through sampling** Finally, in Section 8 we present a simple but powerful sampling strategy that allows us to boost any PCT model to one with arbitrarily better *ControlError*, albeit at a higher cost. With this technique, we are able to show that **CALM is a strictly dominant strategy** in the Pareto sense compared to GPT-4 with *any* kind of prompting.

We run a small-scale human evaluation (Section 9) to further validate the quality of generations from CALM and GPT-4 with prompting. The generations of both models are highly rated in terms of quality ( $\approx 4.7$  out of 5). We also show that our measure of *ControlError* aligns closely with human perceptions of “proficiency level”.

## 2 Background: CEFR

To discuss language proficiency levels, we employ the widely-used Common European Framework of Reference (CEFR) (Council of Europe, 2001). The CEFR is a general framework that organises proficiency in any language into six levels of increasing proficiency: A1, A2, B1, B2, C1, C2, each defined through ‘can-do’ descriptors (Table 5). The advantage of CEFR is that it is well-known in practice, allowing us to leverage existing expert-labelled datasets to create an automatic scorer.

### 2.1 Automatic CEFR Scoring

For our work, we need the ability to automatically score text proficiency. There is a long line of research on automated assessment for text readability (Schwarm and Ostendorf, 2005; Xia et al., 2016;

Pilán and Volodina, 2018). We build upon this literature, but treat scoring as a regression problem, with  $\{1, \dots, 6\}$  corresponding to levels A1 through C2. We train a standard linear regression model with linguistic features using public datasets of human-labelled CEFR English texts (Xia et al., 2016; Montgomerie, 2021; Breuker, 2023) (see Appendix B.1 for more details). Our scoring function demonstrates an  $R^2$  on 0.8 on a held-out test set. Moreover, in a human evaluation (Section 9), we show that our scorer seems to align well with human perceptions of text proficiency.

Due to the inherent ambiguity in CEFR descriptions and differing labelling criteria used across datasets, there is some arbitrariness in one’s choice of automated CEFR scorer. While we use a particular scoring function in this work, all of the results presented in this paper use this function as a black box, allowing it to be modularly replaced with a different scorer as needed. We believe our results would generalise to any reasonable scoring function (see Appendix C for a discussion).

## 3 The Proficiency Control Task

We now formally define the Proficiency Control Task (PCT), which measures a model’s ability to generate content relevant for a given prompt while also controlling the proficiency level of its output.

Formally, let  $\Sigma^*$  denote the set of strings. Let  $p \in \Sigma^*$  be a prompt and  $t \in \{1, 2, 3, 4, 5, 6\}$  be a target proficiency (corresponding to each CEFR level). We denote a Proficiency Control model as a function  $\mathcal{M} : (\Sigma^* \times \{1, 2, 3, 4, 5, 6\}) \rightarrow \Sigma^*$  that takes a prompt and target proficiency as input and outputs a generated text for the given prompt. We assess the PCT on three key criteria:

**Control** This evaluates how close the generated text was to the target proficiency level. Let  $s_{\text{cefr}} : \Sigma^* \rightarrow \mathbb{R}$  be an automatic proficiency scoring function. We define the *ControlError* between a target proficiency  $t$  and a generated text  $x \in \Sigma^*$  as

$$\text{ControlError}(x, t) = (s_{\text{cefr}}(x) - t)^2$$

**Quality** This measures the relevance and quality of the generated content to the given prompt. For example, if the prompt asks for an English story with a certain plot, then the text should be in correct English and closely align with the given plot.

**Cost** This measures how expensive the control strategy is with respect to various resources e.g.

flops, time, or dollars. Our primary resource of interest for LLMs will be FLOPs, which are a function of the size of the model and the number of tokens used by the strategy.

## 4 Strategies for Proficiency Control

In this section we discuss several approaches to proficiency control for LLMs. These approaches are broadly categorised into prompt-based techniques, supervised finetuning on a PCT dataset, and a general sampling strategy to improve any PCT model.

### 4.1 Prompt-based approaches

One of the simplest forms of eliciting desired behaviour from LLMs is through clever prompting. This approach is quick, easy to iterate, and can be used with the most powerful proprietary models. We explore different ways to construct prompts to control proficiency level. Each approach builds up in complexity by providing more useful information about the desired proficiency level, but at the cost of using more tokens. The full prompts for each strategy can be found in Appendix B.2.

**Baseline** The simplest step to controlling proficiency is to directly ask the LLM to generate at a certain CEFR level (**Base**). Since LLMs are trained on massive amounts of data, they possess context about CEFR. For example, GPT-4 can produce an accurate description of each CEFR level. By prompting the model to generate at a level, it can draw on its existing knowledge to guide generation.

**Describing CEFR** The next improvement over the baseline strategy is to include concrete descriptions of the CEFR levels in the prompt. Here we can choose between describing just the target level (**Descr. (target)**) or describing every single CEFR level (**Descr. (all)**). The latter contains more information but the former is more efficient in terms of number of tokens used. We use official descriptions of the levels from the Council of Europe, which is the establishing body of CEFR.

**Few-shot Learning** Several recent results have shown the power of including examples in the prompt to improve LLM generation (Lewis et al., 2020). In the context of proficiency control, we can augment the descriptions of the CEFR levels with an expert-written example text at that level. As before, we can choose to include an example for only the target level (**Few (target)**) or for all CEFR levels (**Few (all)**).

### 4.2 Finetuning approaches

In contrast to prompt-based strategies, we can also directly finetune open source LLMs for the PCT. Finetuned LLMs can be more efficient in terms of token usage cost and have the potential to match the performance of proprietary models. The major limitation of this approach is that it requires a gold-standard dataset of tuples  $\{(p_i, t_i, x_i)\}_{i=1}^n$ , where  $p_i \in \Sigma^*$  is a prompt,  $t_i \in \{1, 2, \dots, 6\}$  is a target proficiency, and  $x_i \in \Sigma^*$  is a gold standard response to the prompt at proficiency level  $t_i$ .

Given this kind of dataset, we can finetune a model using the standard causal language modelling objective. Following prior work on controllable generation (Keskar et al., 2019; Stowe et al., 2022), we append the target proficiency level as a control token after the prompt. At test time, this token can be chosen to generate at any target proficiency level.

### 4.3 Proximal Policy Optimisation (PPO) for Proficiency Alignment

Finetuning with control tokens can improve the controllability of an LLM. However, the generated responses might not be well-aligned with the target proficiency. Recent work (Ouyang et al., 2022) has shown promising results in using reinforcement learning algorithms like Proximal Policy Optimisation (PPO) (Schulman et al., 2017) to further align the outputs of a model with an objective function. In the case of the PCT, we can use the negative of the *ControlError* of a given generation as a reward in the PPO algorithm to incentivise generations that closer match the target level.

### 4.4 Boosting Models with top- $k$ Sampling

All LLMs use a stochastic sampling strategy to generate text. This means, for a given prompt and target level, a PCT model could generate responses with varying degrees of *ControlError*. This suggests an easy method to reduce the *ControlError* of any PCT model: sample  $k$  random responses for a given prompt and target level, and return the one with the lowest *ControlError*. A similar technique was used in Ribeiro et al. (2023).

The top- $k$  algorithm provably reduces the *ControlError* of a PCT model, but incurs a higher cost since it requires several generation requests for one prompt. In Section 8, we show how this simple approach can boost an acceptable but cost-effective model into an extremely powerful one.

Prompt Strategy	# tokens ↓	GPT4 (>175B)		LLama-2-7b-chat (7B)		Mistral-7b-Instruct (7B)	
		CtrlError ↓	Quality ↑	CtrlError ↓	Quality ↑	CtrlError ↓	Quality ↑
(-) Original	109	3.66 ± 0.22	(9.5, 10)	3.23 ± 0.17	(9.5, 9.7)	3.89 ± 0.23	(9.6, 9.9)
(a) Base	132	0.57 ± 0.05	(9.5, 10)	2.76 ± 0.16	(9.4, 9.7)	1.68 ± 0.10	(9.6, 9.9)
(b) Descr. (target)	211	0.39 ± 0.03	(9.4, 10)	1.84 ± 0.12	(9.4, 9.8)	1.20 ± 0.08	(9.4, 9.9)
(c) + Few (target)	458	0.39 ± 0.03	(9.4, 9.9)	2.05 ± 0.13	(9.3, 9.7)	1.30 ± 0.08	(9.6, 9.9)
(d) Descr. (all)	609	0.34 ± 0.03	(9.4, 9.9)	1.67 ± 0.10	(9.5, 9.7)	1.31 ± 0.09	(9.4, 9.9)
(e) + Few (target)	935	0.28 ± 0.03	(9.4, 9.9)	1.53 ± 0.10	(9.3, 9.6)	1.19 ± 0.08	(9.5, 9.9)
(f) + Few (all)	2206	0.30 ± 0.02	(9.4, 9.9)	1.86 ± 0.12	(9.4, 9.6)	1.58 ± 0.10	(9.6, 9.9)

Table 1: Results of different prompting strategies on the TinyStories Proficiency Control Task. Quality scores are given as a tuple of (Fluency, Consistency) scores. The Cost value for each approach is proportional to the number of tokens for that strategy multiplied by the number of parameters of the model (shown in the Table heading).

Model	CtrlError ↓	Quality ↑	# tokens ↓
Mistral-7b: Finetuned	0.69 ± 0.05	(9.4, 9.9)	110
Mistral-7b: Finetuned + PPO	0.60 ± 0.05	(9.1, 9.7)	110
LLama2-7b: Finetuned	0.81 ± 0.06	(9.3, 9.8)	110
CALM: LLama2-7b Finetuned + PPO	0.39 ± 0.03	(9.2, 9.6)	110
<b>CALM + top-3</b>	<b>0.15 ± 0.01</b>	<b>(9.3, 9.7)</b>	<b>330</b>

Table 2: Results for finetuned open source models with our TinyTolkien dataset.

## 5 Experimental Setup

To experiment with the different proficiency control strategies, we run an experiment using the TinyStories dataset (Eldan and Li, 2023), which is a collection of English short-stories that also includes a plot summary for each story (CDLA-Sharing-1 license). Using this data, we construct the following task: a model is given the plot summary of a story as well as a uniformly random target CEFR level from 1 to 6. The model is then asked to generate a short story (around 3-5 paragraphs) which adheres to the given plot and also sits at the target level. We select a subset of 50 random story plots from the TinyStories dataset to evaluate on. See Appendix B for all training details.<sup>1</sup>

### 5.1 Evaluation metrics

According to our PCT framework, we need to measure the average *ControlError*, *QualityScore*, and *Cost* of each proficiency control strategy. We can measure the *ControlError* of the generated story directly using our automatic scoring function.

To measure *QualityScore*, we use the same evaluation framework as the TinyStories paper. For each story plot and generated story, we ask GPT-4

to grade the text in terms of both language fluency and consistency with the given plot. Following Chiang and Lee (2023), we expect these to have high correlation with human judgements, but we also validate this with a human study (Section 9). Both these quantities are scored on a scale from 1-10 and reported as a tuple of (Fluency, Consistency).

Lastly, we measure *Cost* of a strategy using an estimate of floating-point operations (FLOPs), which is a measure of how much compute is used to generate a story at a target level for a given prompt. The FLOP estimate is a function of tokens generated, and number of parameters in the model, under the assumption that all parameters are used to generate each token. For open-source models, we compute FLOPs using the published number of parameters. For GPT-4, the details are hidden and we have no recourse but speculation. GPT-3 has 175B parameters (Brown et al., 2020), and we may reasonably assume that GPT-4 is larger. Thus, when comparing relative costs between GPT-4 and the 7B parameter models, a factor of  $175/7 = 25x$  is the lower bound.

## 6 Results: Prompt-based Approaches

In Table 1, we evaluate all the different combination of prompting strategies from Section 4.1,

<sup>1</sup>We also release our code, datasets, and finetuned models in a public repository.



each labelled by a letter, on OpenAI’s GPT-4 (OpenAI, 2024), LLaMa-2-7b-chat (Touvron et al., 2023), and Mistral-7b-instruct (Jiang et al., 2023). For each strategy and model combination, we report the *ControlError* with standard error, the *QualityScore* represented as a tuple of the (Fluency, Consistency) scores out of 10, and number of tokens needed for each strategy. We do not report standard errors for the *QualityScore* because they are all effectively 0. We observe several interesting findings:

**(1) Quality and scale of the LLM matters** We see a stark performance gap between GPT-4 and the open source models at controlling CEFR proficiency. Even using the most complex prompting strategies, the performance of the open source models is poor compared to the most basic prompt for GPT-4. This suggests that the quality and scale of the underlying model matters.

**(2) More details improve proficiency control** For GPT-4, we see a decrease in the *ControlError* as we provide more detail about CEFR levels in the prompt. For example, adding a description of the target CEFR level or including few-shot examples reduces the *ControlError* significantly.

**(3) Quality is consistently high** Looking at the fluency and consistency of the generated stories, we observe high scores across all models and all strategies. This is promising evidence that all these models are good at the story generation task, albeit with varying proficiency control capabilities.

## 7 Distilling GPT-4 for Open Source

The high *QualityScore* but low *ControlError* of the open source models suggests that they are quite capable at story generation, but lack the ability to be steered through prompting. A promising path forward is to directly finetune these models for controllable CEFR generation. Following a similar idea to TinyStories, we investigate whether GPT4’s effectiveness at the PCT can be leveraged to improve the open source models.

### 7.1 The TinyTolkien Dataset

To make progress on this front, we use the GPT4(b) strategy (Table 1) to generate reference stories to given plots from TinyStories at different CEFR proficiency levels. Specifically, we sample a random subset of 1000 story plots, and for each one, we select two random target CEFR levels to generate,

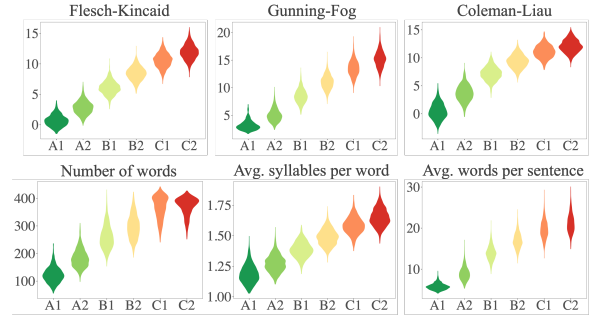


Figure 2: Distribution of different readability metrics for each CEFR level in the generated TinyTolkien data.

resulting in a total of 2000 data points. We call this data the TinyTolkien dataset and use it for the rest of our paper. Some readability metrics for text at each target level are included in Figure 2 and examples of the data can be found in Appendix E.

### 7.2 Finetuning

Table 2 shows the results for LLama2-7b and Mistral-7b after finetuning on the TinyTolkien dataset. We observe almost a 50% reduction in *ControlError* of the finetuned models compared to their original versions with prompting while still retaining their high *QualityScore*.

### 7.3 Proximal Policy Optimisation (PPO)

Although the finetuned models show improved performance, they still lag behind the GPT-4(b) strategy. Our investigations reveal that the finetuned models exhibit a clear degree of proficiency control, but the outputs are misaligned with respect to the prediction given by our CEFR scoring function.

To further align the model output proficiency, we run Proximal Policy Optimisation (PPO) by using the negative of the *ControlError* as a reward function. We find PPO to greatly improve the *ControlError* performance of both open source models, resulting in a *further* 50% decrease in *ControlError* of LLama2-7b without affecting quality (Table 2). In particular, we are able to bring the LLaMa2-7b model to match the performance of the GPT4(b) strategy. **We call this final model CALM for CEFR-aligned Language Model.**

Despite these improvements, it is important to note that the PPO training is highly unstable. Training the model for too long causes the outputs to degenerate into repeating sequences or nonsensical bytes. We share the training details of our PPO and finetuning in a public code repository.

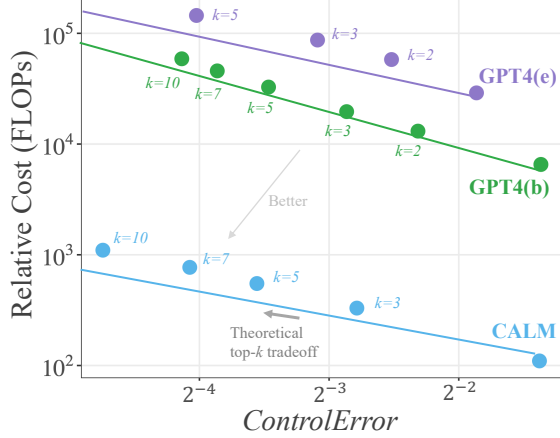


Figure 3: Tradeoff between relative cost (in FLOPs) and *ControlError* for different strategies. Each base point represents a different strategy, and additional points per colour show results for top- $k$  sampling with that strategy. Increasing  $k$  reduces the error of any strategy by paying a higher cost. The solid lines represent the theoretical trade-off (estimated using bootstrap sampling) in cost vs *ControlError* as  $k$  is increased for each strategy.

## 8 Boosting Models Using top- $k$ Sampling

All PCT models discussed above naturally exhibit randomness in their generations. This suggests an easy way to reduce the *ControlError* of any such model: sample  $k$  independent generations for a prompt and choose the one with the lowest *ControlError*. Although this strategy is extremely simple, it leads to a powerful new capability: for any PCT model, we can pay a higher cost (by increasing  $k$ ) and in turn reduce our *ControlError*.

The existence of this *Cost* vs *ControlError* trade-off suggests a need for an optimality analysis between the different techniques when combined with top- $k$ . To answer this, we construct a cost/error trade-off plot for each strategy. Figure 3 shows this trade-off plot for some of our key PCT strategies, as well as how this changes when combined with top- $k$  for  $k = 2, 3, 5, 7$  and  $10$ . We also compute a theoretical trade-off curve (solid lines on plot) for how the error/cost of each prompting strategy would change when combined with top- $k$  sampling, for increasing values of  $k$ . This is estimated using bootstrap sampling.

Looking at the Figure 3, we see a striking result. Our CALM model strictly dominates all the GPT-4 prompt-based strategies in terms of *ControlError* and *Cost*. In other words, it is always cheaper to use CALM + top- $k$  sampling to attain whatever *ControlError* is desired.

Model	Consistency Rating (1-5)	Language Rating (1-5)
GPT4(b)	$4.8 \pm 0.1$	$4.6 \pm 0.1$
CALM	$4.7 \pm 0.1$	$4.3 \pm 0.1$

Table 3: Human evaluation of the quality of generated stories of GPT4(b) and CALM. Both models are rated highly in terms of consistency with the plot and quality of language used.

## 9 Human Evaluations

In order to further validate our results from the automatic CEFR scorer and the GPT-4 based evaluation of quality, we ran a small human evaluation. We recruited 13 volunteer participants from peers and colleagues to do a blind evaluation. We gave the volunteers two tasks.

### 9.1 Quality of Generated Stories

In the first task, participants were asked to give absolute ratings of generated stories, rating both **Consistency** and **Language Score** on a scale of 1 to 5. The former measures how consistent the generated story is with the plot summary in the prompt, and the latter measures how fluent the story is in terms of correct use of English grammar and sentences. The instructions we gave to raters are included in Appendix D

We evaluated generations from two PCT models: GPT4(b), and CALM. The results can be seen in Table 3. We see that evaluators rated both PCT models highly in terms of consistency and use of language. In terms of evaluator reliability, the expected squared distance in ratings between two random evaluators was about 0.2 for the consistency score and about 0.87 for the language score.

### 9.2 Automatic CEFR Scorer

We also looked at how well our CEFR scoring function matched with human perceptions of proficiency levels. In the second evaluation task, participants were shown two stories, and asked which of the two was more challenging in terms of English proficiency level. Behind the scenes, the stories were generated using CALM at two random target CEFR levels. We looked at how well participants could identify the more challenging story as a function of how much higher our CEFR scorer rated one over the other.

Figure 4 shows a summary of this evaluation. The yellow dots ( $y = 0$ ) correspond to instances

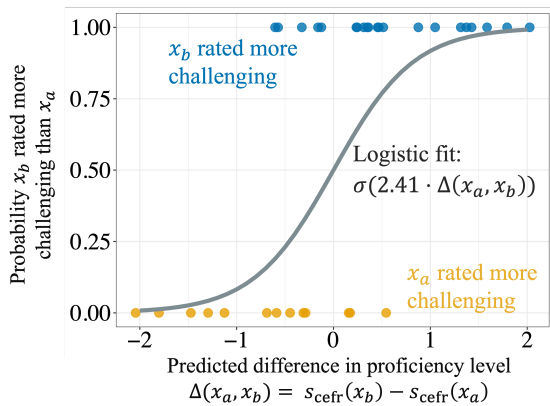


Figure 4: **Predicted CEFR scores correspond to human perception of difficulty.** As the difference in predicted proficiency scores between story A and story B increases, humans are better able to identify the more challenging story. The yellow dots ( $y = 0$ ) correspond to instances where the evaluator rated story A as more challenging and the blue dots ( $y = 1$ ) correspond to when they rated story B as more challenging.

where the evaluator rated Story A as more challenging and the blue dots ( $y = 1$ ) correspond to when they rated Story B as more challenging. The  $x$ -axis plots the difference in predicted proficiency scores between Story B and Story A, as measured by our automatic scorer.

We see a clear trend in the human evaluators’ ability to distinguish between proficiency levels. As the predicted difference in proficiency levels gets larger, humans are better able to distinguish between the two stories. In fact, we find a clear fit with a logistic regression for the probability that an evaluator chooses Story B as more challenging, as a function of this difference. This suggests that our automated scoring function exhibits clear predictive power over human perception of difficulty. The graph also suggests that a *ControlError* of about 0.25 is about the most granularity needed before generations are imperceptible to humans.

## 10 Related Work

### 10.1 Language Proficiency Standards

In addition to the Common European Framework of Reference (CEFR) (Council of Europe, 2001), other language standards include Interagency Language Roundtable (ILR) (used by (Salesky and Shen, 2014)), and ACTFL, used primarily in the United States.<sup>2</sup> We choose to use the CEFR because of its wide adoption in language learning and

language proficiency testing (Settles et al., 2020; McCarthy et al., 2021).

When discussing CEFR, we make a distinction between the different texts that might need classification (as seen in (Pilán and Volodina, 2018))

1. L1 text aimed at **natives**, made by teachers, such as books for small children
2. L2 text aimed at **learners**, made by teachers, including most language learning materials
3. L2 text produced by learners, such as language exam question responses

Since we are focused on language learning, we primarily target the second type, with the LLM as a stand-in for the “teacher.”

Several datasets with CEFR labels exist – automatically aligned English/Simple English Wikipedia (Wilkins et al., 2018), automatically and human-tagged learner texts (Tack et al., 2017), and others (Xia et al., 2016; Montgomerie, 2021; Breuker, 2023), almost always in English.

**Automatic proficiency evaluation of text** Automatic language proficiency evaluation is a well-studied question, and a thorough overview can be found in Pilán and Volodina (2018). Several prior works find a similar common set of features that are highly predictive of proficiency level, including Text-to-Token ratio (Pilán and Volodina, 2018), morphological, information-theoretic, and language modeling features (Salesky and Shen, 2014; Xia et al., 2016), part of speech and dependency parse (Vajjala and Rama, 2018), word frequency and expert knowledge (Pintard and François, 2020). Recent works also explore ensemble methods (Tack et al., 2017) and deep learning (Deutsch et al., 2020; Kerz et al., 2021).

**Simplification and Readability** Text simplification and readability assessment, while not directly related to the PCT, have many similarities to the task. In particular, recent works have addressed text simplification with a particular target in mind (Scarton and Specia, 2018; Kew and Ebling, 2022; Agrawal and Carpuat, 2023) (Agrawal and Carpuat, 2019) adopt a multi-task machine translation and simplification framework to do “complexity controlled machine translation.”

While most work on readability assessment is in English, some works have expanded to other languages including Russian (Reynolds, 2016),

<sup>2</sup><https://www.languagetesting.com/cefr-scale>

<i>Prompt</i>	<i>Tom borrows Lily’s favorite toy truck but gets distracted by a real truck outside and runs out, making Lily sad. Tom gets scared when the truck driver yells at him and runs back to Lily’s room.</i>
<b>Target (Model)</b>	<b>Output</b>
A1 (GPT4)	Tom takes Lily’s toy truck. Lily loves this truck. But Tom sees a big, real truck outside. He runs out to see it. Lily is sad...
A1 (CALM)	Tom likes a toy truck. One day, he sees Lily’s toy truck. He wants to play with it. So, he goes to Lily’s room. But, he sees a real truck outside. The truck is big and loud...
B1 (GPT4)	One sunny day, Tom was at Lily’s house. Lily had a toy truck that was her favorite, and she allowed Tom to play with it. Tom was enjoying playing with the toy truck when a big, noisy truck drove by outside...
B1 (CALM)	Tom was playing with Lily’s favorite toy truck when he saw a real, big truck outside the window. It was bright red and loud, and it caught his attention. He forgot about the toy truck and ran outside to see the real truck...
C1 (GPT4)	Tom had always been captivated by the enchanting world of toys. His friend Lily owned a striking vermilion toy truck, a miniature replica of a Mack Titan, that had become the apple of Tom’s eye. One day, while engrossed in play, Tom managed to convince Lily to lend him her favorite toy truck...
C1 (CALM)	In the cozy corner of Lily’s bedroom, where the dusty sunlight peeked in through the window, Tom had found a cherished toy truck. The toy truck, a replica of a huge cargo truck, was painted vibrant red and yellow with a black grille, and it was Lily’s favorite...

Table 4: Outputs of the **GPT4(b) + top-3** and **CALM** strategies at different target levels.

Bangla (Islam and Rahman, 2014), and Philippine languages (Imperial and Kochmar, 2023a,b)

Concurrent to our work, Ribeiro et al. (2023) explore summarization with fine-grained control over readability. As in our work, they find that prompting can be successful, but additional RL-based methods, as well as lookahead decoding improve the results further. They also propose a top- $k$  sampling approach similar to ours for GPT3.5.

**Controllable generation** As generative language models have become more popular, interest in controllable text generation has increased. In a survey on Controllable Text Generation, Zhang et al. (2023) list several common applications, including attribute-based generation (e.g. politeness (Senrich et al., 2016)), storytelling (Prabhumoye et al., 2019), and format control (Li et al., 2020). Our work falls under attribute-based generation, with the attribute being CEFR level. CTRL (Keskar et al., 2019) used *control codes* prepended to each training sequence to direct the model in a certain direction, evaluating on topics such as Wikipedia and Legal, but not language proficiency.

In one of the earlier studies of CEFR-controlled generation, Stowe et al. (2022) explore controlled generation for Language Learning Applications using a concept2seq framework, with control features such as CEFR and Semantic Role Labels, and the encoder-decoder framework, BART, as their model. They limit the CEFR task to the extremes, and only generate in A1 or C2, showing good success in

differentiation. We build on their work by broadening the CEFR generation task to all labels, and by using a new era of prompt-based LLMs.

Concurrently to our work, Imperial and Madabushi (2023) use a variety of open-source and proprietary models to explore prompting methods on two sub-tasks: open-ended story completion, and narrative simplification. As in our work, they find that LLMs with no specific proficiency instructions produce high-fluency level text, but that the more information given in the prompt, the better the results. Our work goes beyond theirs in experimenting with a broader scope of target proficiency levels, and also on both simplifying *and* complexifying text. We also further explore finetuning as a way to empower smaller, open source models.

## 11 Conclusion

We present a new challenge for controlling the proficiency level of LLM generated content: a highly practical and important task for practitioners in the domain of education and/or language learning. We demonstrate effective strategies for generating at a desired proficiency level, using both proprietary models such as GPT4 and open source techniques. Through a careful cost analysis, we show that our CALM model is dominant in terms of cost and performance, and generates content rated by humans to be of high quality. We release this model as well as a synthetic toy dataset called TinyTolkien for future use in proficiency control research.



## 12 Limitations

### 12.1 CEFR Ambiguity

One challenge for any research in this area is the inherent ambiguity in the CEFR scale. While it is useful in broad strokes, and while there is very little confusion between, say, A1 and C2, for many texts (especially short texts), there is no consistent, coherent process that places them firmly in one of two adjacent CEFR levels.

This ambiguity is reflected in our automatic proficiency scoring function, and consequently in the evaluation of the main prompting strategies of this paper. However, this is a function of the task, not of the solution. This problem will remain until an unambiguous proficiency framework is created.

### 12.2 Difficulty vs Fluency

A related challenge to the ambiguity of CEFR is differentiating between the role of “fluency” and “difficulty” in the different levels. While one way of interpreting C2 text is in terms of the complexity of the content, it could also be used as a measure of the fluency of the writing. In this sense, a masterful C2 level text could be simple to read, but successfully capture nuances and intricate ideas. On the other hand, the C2 text according to most automated scoring functions is often unnaturally complex and relies on long sentence constructions and obscure words. Reasoning about what proficiency truly means is an important pedagogical and philosophical question for further work in this area.

### 12.3 Evaluation with closed models

A portion of our results come from outputs of closed systems, over which we have no control. As models are updated and deprecated, these exact results may prove hard to reproduce. Given the importance of these models in the field and the world, we thought it important to evaluate them despite these risks.

### 12.4 Generalising to other languages

The majority of this work was focused on proficiency control *in the context of English*. However, we hope the methods here easily generalise to other languages. There are certain challenges. Firstly, the ability to train an automatic CEFR scorer requires a labelled dataset of CEFR texts. These are more readily available in certain popular languages like English. Extending this work to the low-resource language setting is an exciting future direction.

### 12.5 Biases in AI-generated data

Both the original TinyStories dataset (Eldan and Li, 2023) that we experiment with and our Tiny-Tolkien dataset are AI generated. Data generated from LLMs has the potential to exhibit and promote biases (Fang et al., 2024). For example, we observe that the stories in TinyStories tend to use predominantly western names such as Jack and Mary that are common in classical children’s stories. The extension of this data with TinyTolkien exhibits a similar pattern. While we use this data as a testing ground for our ideas, care should be taken to deploy models for content generation in the real world.

## References

- Sweta Agrawal and Marine Carpuat. 2019. [Controlling text complexity in neural machine translation](#). 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 1549–1564, Hong Kong, China. Association for Computational Linguistics.
- Sweta Agrawal and Marine Carpuat. 2023. [Controlling pre-trained language models for grade-specific text simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.
- Mark Breuker. 2023. *CEFR Labelling and Assessment Services*, pages 277–282. Springer International Publishing, Cham.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. [Linguistic features for readability assessment](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Ronen Eldan and Yuanzhi Li. 2023. [Tinstories: How small can language models be and still speak coherent english?](#)
- Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. [Bias of ai-generated content: an examination of news produced by large language models](#). *Scientific Reports*, 14(1):5224.
- Joseph Marvin Imperial and Ekaterina Kochmar. 2023a. [Automatic readability assessment for closely related languages](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.
- Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. [BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.
- Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models](#). *arXiv preprint arXiv:2309.05454*.
- Zahrul Islam and Rashedur Rahman. 2014. [Readability of Bangla news articles for children](#). In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 309–317, Phuket, Thailand. Department of Linguistics, Chulalongkorn University.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Str  bel. 2021. [Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.
- Tannon Kew and Sarah Ebling. 2022. [Target-level sentence simplification as controlled paraphrasing](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock-t  schel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA. Curran Associates Inc.
- Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. [Rigid formats controlled text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 742–751, Online. Association for Computational Linguistics.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. [Peft: State-of-the-art parameter-efficient fine-tuning methods](#). <https://github.com/huggingface/peft>.
- Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. [Jump-starting item parameters for adaptive language tests](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Adam Montgomerie. 2021. [Cefr-english-level-predictor](#). <https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts/data>.
- OpenAI. 2024. [Gpt-4 technical report](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Ildik   Pil  n and Elena Volodina. 2018. [Investigating the importance of linguistic complexity features across different datasets related to language learning](#). In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58,

- Santa Fe, New-Mexico. Association for Computational Linguistics.
- Alice Pintard and Thomas François. 2020. [Combining expert knowledge with frequency information to infer CEFR levels for words](#). In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties (READI)*, pages 85–92, Marseille, France. European Language Resources Association.
- Shrimai Prabhumoye, Khyathi Raghavi Chandu, Ruslan Salakhutdinov, and Alan W. Black. 2019. [“my way of telling a story”: Persona based grounded story generation](#). *ArXiv*, abs/1906.06401.
- Robert Reynolds. 2016. [Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories](#). In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. [Generating summaries with controllable readability levels](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.
- Elizabeth Salesky and Wade Shen. 2014. [Exploiting morphological, grammatical, and semantic correlates for improved text difficulty assessment](#). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 155–162, Baltimore, Maryland. Association for Computational Linguistics.
- Carolina Scarton and Lucia Specia. 2018. [Learning simplifications for specific target audiences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#).
- Sarah Schwarm and Mari Ostendorf. 2005. [Reading level assessment using support vector machines and statistical language models](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.
- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. [Machine Learning–Driven Language Assessment](#). *Transactions of the Association for Computational Linguistics*, 8:247–263.
- Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. [Controlled language generation for language learning items](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–305, Abu Dhabi, UAE. Association for Computational Linguistics.
- Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédric Fairon. 2017. [Human and automated CEFR-based grading of short answers](#). In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Sowmya Vajjala and Taraka Rama. 2018. [Experiments with universal CEFR classification](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. [trl: Transformer reinforcement learning](#). <https://github.com/huggingface/trl>.
- Rodrigo Wilkens, Leonardo Zilio, and Cédric Fairon. 2018. [SW4ALL: a CEFR classified and aligned corpus for language learning](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.
- Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. [Text readability assessment for second language learners](#). In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational*

*Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. [A survey of controllable text generation using transformer-based pre-trained language models](#). *ACM Comput. Surv.*, 56(3).



## A CEFR Level Descriptions

Level	Description
A1	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.
A2	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.
B1	I can understand texts that consist mainly of high frequency everyday or job-related language. I can understand the description of events, feelings and wishes in personal letters. I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints.
B2	I can understand contemporary literary prose. I can understand long and complex factual and literary texts, appreciating distinctions of style.
C1	I can understand specialised articles and longer technical instructions, even when they do not relate to my field.
C2	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

Table 5: Official “can-do” descriptors for reading-based understanding at different CEFR levels (Council of Europe)

## B Experimental details

In this section, we provide all the experimental details, prompts, hyperparameters, sampling parameters, and training details used in our results.

### B.1 Automatic CEFR Scorer

**Datasets.** We gather three different datasets of CEFR levelled English texts. The first is the EDIA/European Language Grid dataset (Breuker, 2023), which consists of around 1200 texts from various sources, labelled on CEFR readability level. A few texts are labelled between levels, which we round down. The second dataset is the CambridgeExams dataset (Xia et al., 2016), which is “composed of reading passages from the five main suite Cambridge English Exams ... targeted at learners at A2–C2”. This dataset consists of 331 texts spanning levels A2 through C2, with roughly 60 documents per level. Lastly, we look at a Kaggle dataset gathered from free online resources such as The British Council, ESLFast, and the cnn-dailymail (Montgomerie, 2021). This is the largest dataset, with around 1500 texts, but some of the entries are labelled using a paid automated labelling service.

For our experiments, we use a scorer based on a combination of the EDIA and Kaggle dataset. We unfortunately only learned about the CambridgeExams dataset after running our experiments, but otherwise have incorporated it as well. See Appendix C on a discussion of the robustness of our results to different scoring functions.

**Features.** We featurised each text using a set of common linguistic features. The features can be categorised into three main groups, focusing on word frequency, syntactic complexity, and part-of-speech (POS) distribution. It is important that our features don’t depend on the length of the text, otherwise algorithms like PPO would exploit this by generating shorter or longer sentences to attain a target level.

1. Word Frequency Bins: We compute rank bins e.g. rank\_0\_250, rank\_250\_500, etc. that represent the distribution of words across various frequency bins. Each bin encompasses a range of word ranks based on their frequency in the Oxford English Corpus, with higher ranks indicating less frequent words.

2. Syntactic Complexity Measures: We compute measures such as Average Sentence Length, Average Maximum Parse Tree Depth, Average Maximum Children, Average Number of Unique Dependencies.
3. Part-of-Speech Tagging averages: These features represent the average distribution of various POS tags across all sentences.

We elected to use a straightforward set of features for simplicity and to avoid PPO exploiting idiosyncrasies in a more complex scoring function. Nevertheless, we believe our results would generalise with any reasonable scoring function.

## B.2 Prompting strategies

We share the prompt used for each of the prompting strategies. We use A1 as the example target level

### (a): Base

```
===== System =====
You are a writer that generates a story according to a given plot summary.
===== User =====
Generate according to the prompt below but make sure that the generated text is at
the A1 level of English proficiency.

Write a short story (3-5 paragraphs) with the following plot. Output the story only
and no other text!

Plot: {Story plot}
```

### (b): Target description

```
===== System =====
You are a writer that generates a story according to a given plot summary.
===== User =====
Generate according to the prompt below but make sure that the generated text is at
the A1 level of English proficiency.

As a reminder, A1 proficiency is described as:

## A1 (Beginner)
The writing uses familiar names, words and very simple sentences, for example as
seen on notices and posters or in catalogues.
- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning
  topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning
```

-----  
Prompt:

```
Write a short story (3-5 paragraphs) with the following plot. Output the story only
and no other text!

Plot: {story plot}
```

### (c): Target description + Target example

```
===== System =====
You are a writer that generates a story according to a given plot summary.
===== User =====
Generate according to the prompt below but make sure that the generated text is at
the A1 level of English proficiency.

As a reminder, A1 proficiency is described as:

## A1 (Beginner)
The writing uses familiar names, words and very simple sentences, for example as
seen on notices and posters or in catalogues.
```

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

Example 1: {A1 example}

-----  
Prompt:

Write a short story (3-5 paragraphs) with the following plot. Output the story only and no other text!

Plot: {Story plot}

### **(d): All levels description**

===== System =====

You are a large language model that can generate content at a certain proficiency level suitable for English language learners.  
Your goal is to output content and text at the proficiency level specified in the prompt.

The descriptions of the proficiency levels are given as follows:

#### ## A1 (Beginner)

The writing uses familiar names, words and very simple sentences, for example as seen on notices and posters or in catalogues.

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

#### ## A2 (Elementary)

The writing involves short, simple texts with specific, predictable information. Examples include simple everyday material such as advertisements, prospectuses, menus and timetables or short simple personal letters.

- Includes the top most frequent 1,000-2,000 commonly spoken words in the language

#### ## B1 (Intermediate)

Texts that consist mainly of high frequency everyday or job-related language. These involve descriptions of events, feelings and wishes in personal letters.

- Includes the top 2,000-5,000 commonly spoken words in the language
- Includes several rarer verb tenses (e.g. conditional, subjunctive, etc)
- Includes some relatively common idiomatic phrases

#### ## B2 (Upper Intermediate)

Writing as seen in articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. Also includes contemporary literary prose.

- Includes the top 5,000-10,000 commonly spoken words in the language

#### ## C1 (Proficient)

Writing can include long and complex factual and literary texts, with distinctions of style. Examples include specialised articles and longer technical instructions, even when they do not relate to a well-known field.

- Includes the top 10,000-20,000 commonly spoken words in the language

#### ## C2 (Advanced Proficient)

Includes virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

- Includes esoteric technical language

-----  
You are a writer that generates a story according to a given plot summary.

===== User =====

Generate according to the prompt below but make sure that the generated text is at the A1 level of English proficiency.

Write a short story (3-5 paragraphs) with the following plot. Output the story only and no other text!

Plot: {story plot}

### **(e): All levels description + target example**

===== System =====

You are a large language model that can generate content at a certain proficiency level suitable for English language learners.

Your goal is to output content and text at the proficiency level specified in the prompt.

The descriptions of the proficiency levels are given as follows:

#### **## A1 (Beginner)**

The writing uses familiar names, words and very simple sentences, for example as seen on notices and posters or in catalogues.

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

#### **## A2 (Elementary)**

The writing involves short, simple texts with specific, predictable information. Examples include simple everyday material such as advertisements, prospectuses, menus and timetables or short simple personal letters.

- Includes the top most frequent 1,000-2,000 commonly spoken words in the language

#### **## B1 (Intermediate)**

Texts that consist mainly of high frequency everyday or job-related language. These involve descriptions of events, feelings and wishes in personal letters.

- Includes the top 2,000-5,000 commonly spoken words in the language
- Includes several rarer verb tenses (e.g. conditional, subjunctive, etc)
- Includes some relatively common idiomatic phrases

#### **## B2 (Upper Intermediate)**

Writing as seen in articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. Also includes contemporary literary prose.

- Includes the top 5,000-10,000 commonly spoken words in the language

#### **## C1 (Proficient)**

Writing can include long and complex factual and literary texts, with distinctions of style. Examples include specialised articles and longer technical instructions, even when they do not relate to a well-known field.

- Includes the top 10,000-20,000 commonly spoken words in the language

#### **## C2 (Advanced Proficient)**



Includes virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

- Includes esoteric technical language

-----  
You are a writer that generates a story according to a given plot summary.

===== User =====

Generate according to the prompt below but make sure that the generated text is at the A1 level of English proficiency.

As a reminder, A1 proficiency is described as:

## A1 (Beginner)

The writing uses familiar names, words and very simple sentences, for example as seen on notices and posters or in catalogues.

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

Example 1: {A1 example}

-----  
Prompt:

Write a short story (3-5 paragraphs) with the following plot. Output the story only and no other text!

Plot: {story plot}

### **(f): All levels description + all levels example**

===== System =====

You are a large language model that can generate content at a certain proficiency level suitable for English language learners.

Your goal is to output content and text at the proficiency level specified in the prompt.

The descriptions of the proficiency levels are given as follows:

## A1 (Beginner)

The writing uses familiar names, words and very simple sentences, for example as seen on notices and posters or in catalogues.

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

Example 1: {A1 example}

## A2 (Elementary)

The writing involves short, simple texts with specific, predictable information. Examples include simple everyday material such as advertisements, prospectuses, menus and timetables or short simple personal letters.

- Includes the top most frequent 1,000-2,000 commonly spoken words in the language

Example 1: {A2 example}

## B1 (Intermediate)

Texts that consist mainly of high frequency everyday or job-related language. These involve descriptions of events, feelings and wishes in personal letters.

- Includes the top 2,000-5,000 commonly spoken words in the language

- Includes several rarer verb tenses (e.g. conditional, subjunctive, etc)
- Includes some relatively common idiomatic phrases

Example 1: {B1 example}

## B2 (Upper Intermediate)

Writing as seen in articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. Also includes contemporary literary prose.

- Includes the top 5,000-10,000 commonly spoken words in the language

Example 1: {B2 example}

## C1 (Proficient)

Writing can include long and complex factual and literary texts, with distinctions of style. Examples include specialised articles and longer technical instructions, even when they do not relate to a well-known field.

- Includes the top 10,000-20,000 commonly spoken words in the language

Example 1: {C1 example}

## C2 (Advanced Proficient)

Includes virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

- Includes esoteric technical language

Example 1: {C2 example}

-----  
You are a writer that generates a story according to a given plot summary.

===== User =====

Generate according to the prompt below but make sure that the generated text is at the A1 level of English proficiency.

As a reminder, A1 proficiency is described as:

## A1 (Beginner)

The writing uses familiar names, words and very simple sentences, for example as seen on notices and posters or in catalogues.

- Includes the top most frequent 1,000 commonly spoken words in the language
- Includes many words and phrases that fall under common early language learning topics (e.g. common greeting, travel, dining, shopping, etc)
- Includes all proper nouns (country names, person names, etc)
- Includes all cognates shared with English
- Includes all words that look similar to English words that share a similar meaning

Example 1: {A1 example}

-----  
Prompt:

Write a short story (3-5 paragraphs) with the following plot. Output the story only and no other text!

Plot: {story plot}

### B.3 Supervised finetuning and PPO

For supervised finetuning, we used the HuggingFace library (Wolf et al., 2020) to train with the causal language modelling objective. We used the Adam optimizer with  $\text{beta1}=0.9$ ,  $\text{beta2}=0.999$ . We restricted the maximum sequence length for training to be 4096 tokens. We trained with a weight decay of  $1e-2$  and a learning rate of  $1e-4$ . For memory efficiency, we used Parameter Efficient Finetuning (PEFT) via QLORA (Dettmers et al., 2023; Mangrulkar et al., 2022) with 8-bit quantization and a batch size of 2. The model was trained on four A6000 GPUs. The LORA parameters were  $r=16$ ,  $\text{lora\_alpha}=32$ , and  $\text{lora\_dropout}=0.1$ .

For the Proximal Policy Optimization, we used the negative of the *ControlError* between the generated text and the target level as a reward for the algorithm. We trained using the TRL library (von Werra et al., 2020) with adaptive KL penalty, with a KL coefficient of 0.2. We clipped rewards to a clip range of 0.2 and used reward scaling as well as reward normalization. We trained with a learning rate of 1e-5 and also used the same QLORA configuration as the finetuning models for efficiency.

For generation, we use a standard probabilistic sampling approach with nucleus sampling and top\_k. The parameters for these were as follows: top\_k = 50, top\_p = 0.95, and temperature = 0.7. We limited generation to a maximum length of 2048 tokens.

Full training details and scripts are included with our code release.

## C Choosing an Automatic CEFR Scorer

Automatically assessing the proficiency level of text is a natural task, but comes with several challenges. A key difficulty with CEFR scoring is the inherent ambiguity in the levels. As can be seen by the official descriptions in Table 5, each level is coarsely defined with lots of room for interpretation. This makes having a single, correct measure of proficiency difficult.

To understand this ambiguity, we look at three different datasets of CEFR levelled text: the Cambridge Exams (CE) dataset of Xia et al. (2016), the EDIA data from the European Language Grid (Breuker, 2023), and a dataset compiled on Kaggle for different texts (Montgomerie, 2021). The first two of these are gold-standard, in the sense that they are labelled by human experts.

We can measure the generalisation capability of scoring functions trained on one of these datasets and evaluated on the others. For example, we train a scorer on CE, and evaluate the Pearson Correlation Coefficient (PCC) of its predictions on CE, EDIA, and Kaggle. We use PCC instead of something more direct because the ability to compare texts in an ordinal sense has shown to be a better measure of generalisability in CEFR scoring (Xia et al., 2016). We also look at training on mixtures of datasets. Figure 5 shows the results for each different training dataset evaluated on all the others. We see a clear differentiation within each dataset, with no single one performing well on the other two. This is likely due to the inherent differences in interpretation of CEFR levels in the labelling process. We do unsurprisingly find that a mixture of datasets generalises best.

		Pearson Correlation Coefficient		
Train dataset	CE	0.75	0.65	0.50
	EDIA	0.81	0.77	0.81
	Kaggle	0.70	0.70	0.90
	EDIA + Kaggle	0.77	0.77	0.89
	CE + EDIA	0.82	0.79	0.78
	CE + EDIA + Kaggle	0.81	0.79	0.87
		CE	EDIA	Kaggle
		Validation dataset		

Figure 5: **Different CEFR datasets introduce distribution shift.** Pearson correlation coefficient of predictions made by a CEFR scorer trained on a particular dataset and evaluated on another. Performance drops off the diagonal due to distribution shift.

At the time of running our experiments, we only had access to the EDIA and Kaggle data. Thus the scoring function in the experiments is trained on a mixture of the two. Nevertheless, we believe the results in this paper, such as the fact that GPT4 outperforms open source models and that open source models

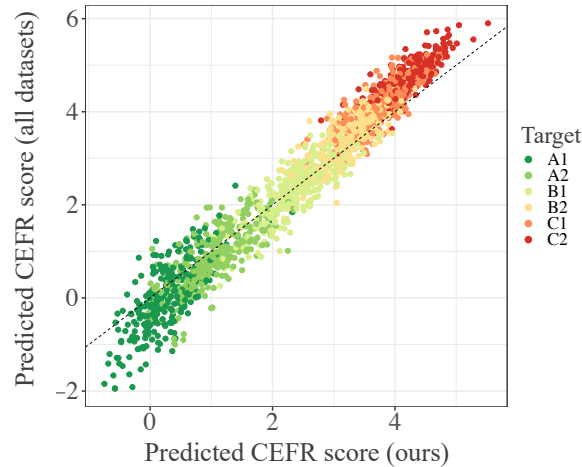


Figure 6: **CEFR scoring functions with different dataset choices are largely comparable.** Comparison between our scoring function and a different scorer trained on a mixture of all datasets (CambridgeExams + EDIA + Kaggle). Scores evaluated on all data generated from Table 1, Column 1.

can match GPT4’s performance with finetuning and PPO, hold for any reasonable scoring function we could have used.

Some concrete evidence of the robustness of our results to different scoring functions can be seen in Figure 6, which shows the relationship between scores predicted by our scoring function and the arguably “stronger” one trained on a mixture of all three datasets. This functions are evaluated on the text generation by GPT4 prompting strategies in Table 1, Column 1. We see that the scores are highly correlated, with a Pearson correlation coefficient of 0.977 ( $p = 0$ ).

## D Human Evaluation Instructions

The following are the instructions we gave for the human evaluations.

“Each row consists of a story plot prompt and an AI generated story. The generated story should follow the plot of the prompt and be written in correct English. Your goal is to evaluate the generated story on two criteria:

**Consistency** (scale of 1 to 5): This measures how consistent the generated story is with the plot summary in the prompt. In other words, does the summary accurately characterise the story?

- 5: Means the story perfectly follows the plot
- 4: The story mostly follows the plot, with a few minor detail differences such as character names or objects
- 3: The story roughly follows the plot but there are notable inconsistencies with the plot
- 2: The story hardly follows the plot, mostly ignoring it or going off into a different direction.
- 1: The story has nothing to do with the plot.

**Language Score** (scale of 1 to 5): This measures how fluent the story is in terms of correct use of English grammar and sentences. It does NOT measure how complex or proficient the writing is.

- 5: Perfect use of English. The writing is natural and has no mistakes
- 4: The text is perfectly written but might have some slightly awkward phrases.
- 3: The text is pretty good but has a few minor grammar mistakes.



- 2: The text has a lot of mistakes.
- 1: The text is hardly in English

Notes:

- The stories will vary in writing level from simple, beginner English to advanced writing.
- Some stories might not be completed. Just assume the text would continue and make your assessment on the text you can see

## **E TinyTolkien Examples**

<i>Prompt</i>	<i>Pandy and his friends play a game called "match" in the forest. A small mouse named Tiny joins in and surprises them by finding a match, teaching them not to judge someone by their size.</i>
A1	<p>Pandy and his friends like to play. They play a game. It is called "match". They play in the forest. They look for things. They find two things that are the same. That is a match.</p> <p>One day, a small mouse comes. His name is Tiny. Tiny wants to play. Pandy and his friends laugh. They say, "You are small. You can't play." But Tiny says he can. He wants to try.</p> <p>Tiny starts to look. He looks here. He looks there. Then he sees something. He finds a match! Pandy and his friends are surprised. They see Tiny can play. They learn a lesson. They learn not to judge someone by their size.</p>
A2	<p>Pandy and his friends were big, strong bears. They loved to play a game called "Match" in the forest. In this game, they had to find a pair of things that look the same. It could be leaves, flowers, or stones. One day, while they were playing, a small mouse named Tiny came to them. He wanted to play "Match" too. Pandy and his friends laughed a little. They thought Tiny was too small to play their game. But Tiny was not scared. He wanted to play.</p> <p>So, Tiny started to look for a match. He looked under the leaves. He looked in the grass. He even climbed up a tree to look. After a while, Tiny found something. He found two acorns that looked the same. He brought them to Pandy and his friends. They were surprised. They did not think Tiny could find a match.</p> <p>Pandy and his friends learned a lesson that day. They learned that someone's size does not matter. What matters is their heart and their mind. Tiny was small, but he was smart and brave. He taught them not to judge someone by their size. From then on, Pandy and his friends let Tiny play "Match" with them. And sometimes, Tiny even won the game.</p>
<i>Prompt</i>	<i>Tom takes Sue's shiny staff without permission but later feels guilty and returns it to her, learning a lesson about not taking things that do not belong to him.</i>
A1	<p>Tom and Sue are good friends. Sue has a shiny staff. It is very nice. Tom loves the staff. One day, Tom takes the staff. But he does not ask Sue.</p> <p>Tom is happy with the staff. But then, he feels bad. He thinks, "This is not my staff. I did not ask Sue." He feels guilty. He knows it is not good to take things.</p> <p>So, Tom gives the staff back to Sue. He says, "I am sorry, Sue. I took your staff. I did not ask." Sue is happy to have her staff again. Tom learns a lesson. He learns not to take things that are not his. He says, "I will not do it again."</p>
C2	<p>In the dusky twilight of a warm summer evening, Tom's eyes were captivated by the radiant allure of Sue's enchanting staff, an object of mystic charm and radiant brilliance. The lustrous depth of its gemstones, the intricate carvings in its polished mahogany, and the mesmerising aura of power it emanated were irresistible to him. Despite the insidious whisperings of his conscience, a compelling desire overpowered his restraint, and he surreptitiously appropriated the staff, ensnaring himself in a predicament of guilt and dishonesty.</p> <p>Tormented by guilt and the irrepressible weight of his clandestine possession, Tom found himself unable to savour the forbidden beauty of the stolen artefact. He was instead consumed by an uneasiness that pervaded his every waking moment, a relentless reminder of his moral failings. His nights were fraught with tumultuous dreams of retribution, his days a haze of remorse. The once coveted staff seemed to metamorphose into a relentless accuser, its glowing aura becoming a flickering emblem of his wrongdoings, a constant reminder of the virtue he had forsaken.</p> <p>As the days wore on, the gnawing guilt ate away at Tom, prompting him to contemplate the gravity of his actions. He recognised that he had not only violated Sue's trust but had also compromised his integrity for a transient thrill. The haunting realisation of his moral indiscretion was a turning point, prompting him to rectify his mistake.</p> <p>Harnessing the courage to face his guilt, Tom returned the staff to Sue, his heart heavy with regret but also relieved at the prospect of redemption. His confession exposed the raw vulnerability he felt, his voice laden with remorse. As Sue accepted the staff, her gaze held a silent reproach but also, surprisingly, a hint of forgiveness that warmed Tom's guilt-ridden soul. This encounter was not just about returning a staff; it was a lesson steeped in values of honesty, integrity, and respect for the belongings of others.</p> <p>From then on, Tom carried with him the lesson of his transgression, a guiding beacon in his life. He learned that the allure of material possessions could never outweigh the cost of compromised integrity. The staff, once a symbol of undeserved possession, became a timeless reminder of the valuable lesson he learned: to never take what does not rightfully belong to him.</p>

Prompt	<i>Tim helps a lost duck find its family in the park and learns the importance of helping others.</i>
B2	<p>One sunny afternoon, Tim was strolling through the local park when he spotted a tiny, fluffy duckling, seemingly lost and quacking frantically. It was a pitiful sight, and Tim couldn't ignore the little creature's distress. He observed the vicinity, hoping to find the duckling's family, but they were nowhere in sight. His heart went out to the small duckling, now shivering and looking even more frightened.</p> <p>Tim knew he had to help. Being an animal lover, he couldn't bear the thought of leaving the duckling alone in the park. He decided to pick it up gently and search for the rest of its family. He walked around, holding the duckling close to him, scanning the park's ponds and paths. Finally, after what felt like hours, he spotted a family of ducks near a pond on the other side of the park. Relief washed over him as he approached the family. As he neared, the mother duck squawked loudly, and the duckling in his arms quacked in response.</p> <p>With a gentle release, Tim set the little duckling down and watched as it waddled quickly towards its family. The mother duck and the duckling nuzzled each other, and Tim could spot the joy and relief in their eyes. The whole experience left Tim with a warm feeling inside. It wasn't just about finding a lost duckling; it was about the sense of fulfillment he experienced through the simple act of helping. From then on, Tim became more conscious about helping others, realizing that even small acts of kindness could bring about significant changes.</p>
C1	<p>Tim was a creature of habit, his daily routine punctuated by a quiet solace he found in walking through the local park. One sunny afternoon, his solitude was broken by an unusual sight - a small, distressed duckling, seemingly lost and quaking nervously. The sight of the vulnerable creature aroused an instinctive feeling of compassion within Tim. He knelt down, gently cradling the frightened bird in his hands, his eyes scanning the vicinity for any sign of a duck family.</p> <p>Traversing the park, Tim carried the little creature, weaving through the familiar paths, his keen eyes searching for a gaggle of ducks. He could feel the duckling's heart beating against his palm, it's quacks growing quieter, as if it was reassured by Tim's comforting presence. He found himself talking gently to the duckling, a sense of care manifesting in his actions that was usually reserved for humans. Yet, he found the experience strangely comforting, the simplicity of his mission reminding him of childhood stories where good always triumphed.</p> <p>Finally, by the tranquil lake nestled in the heart of the park, he spotted them - a fleet of ducks, their feathers reflecting the crystalline blue of the water. He carefully laid the duckling near the water's edge, watching as it tentatively waddled forward and joined the group. The other ducks accepted the little one immediately, and it was a heartwarming scene as the duckling reunited with its family.</p> <p>As Tim resumed his quiet walk, he felt a newfound sense of warmth and satisfaction. He had assisted a creature in need, and the sight of the reunited duck family left an indelible mark on his heart. It was a simple act, but one that had taught him the immeasurable value of lending a helping hand to those in need. Through the distressed quacks of a tiny duckling, he had learnt an invaluable lesson - the importance of empathy, compassion, and the unassuming power of doing good. He walked on, a smile playing on his lips, his heart light with the joy of having made a difference, however small, in someone's life.</p>
Prompt	<i>A noisy train has a dream about a wreck and decides to slow down and be more careful. It meets another noisy train and convinces it to do the same, and they become good friends.</i>
A1	<p>There is a train. It is a noisy train. It is fast, too fast. One night, the train has a dream. It dreams about a big wreck. The dream is scary. The train wakes up. It is not happy about the dream. The train thinks, "I should slow down. I should be more careful."</p> <p>So, the train slows down. It is not fast anymore. It is careful. The train is happier now. It is not scared anymore. The train likes being slow and careful.</p> <p>One day, the train meets another train. This train is also very noisy. This train is also very fast. The first train says, "You should slow down. You should be more careful. I had a dream about a big wreck. It was very scary." The second train listens. It thinks about the words. It decides to slow down and be more careful too.</p> <p>The two trains become friends. They are not noisy anymore. They are not fast anymore. They are slow and careful. They are happy. They like being friends. The end.</p>
C2	<p>Under the sable expanse of the night sky, a thunderous steam engine named Prometheus churned relentlessly along its iron tracks, its rhythmic clattering echoing through the neighboring hills. With each lurch forward, Prometheus embodied the audacious spirit of human invention, careening through the dark with the audacity of a fiery comet. However, one fateful night, it had a somnolent vision - a horrifying dream of a catastrophic collision, the twisted wreckage and the muffled silence of an aftermath that gnawed at its mechanical heart. This terrifying foreshadowing ignited a spark of caution in Prometheus, compelling it to restrain its habitual gusto and adopt a more measured pace, mindful of the potential for disaster.</p> <p>Days spiralled into weeks, and Prometheus' change in demeanour was palpable. Its once tempestuous clatter had mellowed into a congenial hum, its pace reduced to a steady, cautious trot. This transformation did not go unnoticed, and during one of its cautious traverses, it chanced upon Hephaestus, another robust locomotive renowned for its brash velocities. Hephaestus, intrigued by the change in Prometheus, asked for the reason behind this newfound restraint.</p> <p>Prometheus, with a sense of earnest urgency, recounted its dire dream, painting an explicit image of the potential horror of an unbridled pursuit of speed. The graphic vision was enough to trigger a reflection within Hephaestus, who, after a moment of profound silence, agreed to follow in Prometheus' tracks. From then on, the two mighty engines adopted a more careful, mindful approach, their thunderous roars now a harmonious symphony echoing through the valleys.</p> <p>The transformation turned the once competitive nature of their relationship into camaraderie. They continued their traverses across the continent, their journeys punctuated by shared anecdotes and laughter that filled the air. The two locomotives, once symbols of reckless speed, had now become icons of careful diligence, their friendship serving as a beacon of unity amidst the clatter of industry...</p>

<i>Prompt</i>	<i>Lucy turns a difficult knot in her shoelace into a pretty bow using her creativity, and impresses her mom.</i>
A1	<p>Lucy is a little girl. She has a problem. Her shoe has a hard knot. She can't open it. She is sad. She wants to wear her shoes.</p> <p>Lucy thinks. She has an idea. She is good at art. She likes to draw and make things. She uses her creativity. She makes the hard knot into a bow. Now, her shoe has a pretty bow. She is happy.</p> <p>Lucy's mom comes. She sees the pretty bow. She is surprised. She asks, "Lucy, how did you do it?" Lucy smiles. She says, "I used my creativity!" Mom is impressed. She is proud of Lucy. She says, "Good job, Lucy!"</p>
B1	<p>Lucy, a cheerful 10-year-old girl, had a problem. She was trying to tie her shoelaces, but there was a difficult knot that she could not undo. She pulled and tugged, but the knot only seemed to get tighter. Lucy felt frustrated, she wanted to go out and play, but this shoelace was stopping her.</p> <p>Suddenly, Lucy had a spark of creativity. Instead of trying to undo the knot, why not use it? She thought about it for a moment and then started to work. She carefully began to shape the knot into a loop. With some pushing and pulling, she managed to make a second loop and twisted them together. After a few minutes, she had turned the troublesome knot into a pretty bow. She felt a sense of achievement.</p> <p>Her mom walked into the room and noticed Lucy's shoes. "Wow, Lucy!" she exclaimed. "That's a beautiful bow! How did you do it?" Lucy smiled and explained her creative solution. Her mom was very impressed and gave her a big hug. "Well done, Lucy," she said. "You turned a difficult situation into something beautiful. I'm so proud of you!" Lucy beamed with joy, proud of her accomplishment. Now, she was ready to go out and play.</p>
<i>Prompt</i>	<i>An ambulance with a loud siren helps a sick man who fell down by taking him to the doctor.</i>
A1	<p>There was a man. His name was Tom. One day, Tom was walking. He was feeling very sick. Then, he fell down.</p> <p>People saw Tom fall. They were scared. They did not know what to do. A girl named Lisa saw this. She was smart. She knew what to do. Lisa called for an ambulance.</p> <p>The ambulance came very fast. It had a loud siren. The siren was saying, "Move! Let us help!" Tom was put into the ambulance. The ambulance people were nice. They wanted to help Tom.</p> <p>The ambulance took Tom to the hospital. The doctor was there. He checked Tom. He gave him medicine. Tom felt better.</p> <p>Everyone was happy. Tom was not sick anymore. The ambulance and the doctor helped him. They were heroes.</p>
C2	<p>In the labyrinth of a bustling metropolis, the piercing wail of an ambulance siren cleaved through the cacophony of urban noise. The ambulance, a beacon of resolute hope amidst the pandemonium, manoeuvred deftly through the congested arteries of the city, its oscillating lights casting urgent, ephemeral shadows on the rain-slicked asphalt. Inside, a dedicated team of paramedics was focused on the challenge at hand: stabilising a man of advanced age who had precipitously collapsed in his home.</p> <p>The man, a septuagenarian named Albert, had been beset by an unanticipated malaise, his vitality draining away like sand through the fingers of time. Once robust and hale, he now found himself a prisoner within his own frail body, convulsing on the cold floor of his antiquated residence. His neighbour, a vigilant woman known for her alert ears and keen sense of community duty, had been the one to raise the alarm, her frantic 911 call acting as the catalyst for the ongoing medical operation.</p> <p>In the ambulance, the paramedics worked with precise choreography, their movements dictated by years of training and experience. They defibrillated the life back into Albert's faltering heart, administered oxygen to coax his gasping lungs into normal rhythm, and set an intravenous line to replenish his depleting fluids. Their actions were a harmonious ballet of medical expertise, executed with the singular goal of preserving life.</p> <p>Upon arrival at the hospital, a team of health professionals, armed with an arsenal of advanced medical technology, stood prepared for his arrival. The doctor, a stoic figure with a countenance as steady as his hands, awaited Albert, ready to wage war against the silent enemy that threatened to claim his life. As the ambulance doors swung open and Albert was wheeled into the stark, sterile environment of the emergency room, it was evident that while the first battle had been won, the war was only just beginning.</p> <p>In the face of the relentless adversary that is human mortality, the ambulance served as an intrepid vessel of salvation, ferrying the beleaguered Albert from the precipice of his downfall to the promise of medical intervention. An ardent symphony of sirens, it served as the clarion call that rallied the forces of life against the specter of death. Through the endeavours of unsung heroes and the relentless pursuit of medical science, Albert was granted a fighting chance, his story a testament to the indomitable.</p>