

GANDALF: a General Character Name Description Dataset for Long Fiction

Fredrik Carlsson¹, Fredrik Olsson², Amaru Cuba Gyllensten¹, and Magnus Sahlgren¹

¹RISE, Isafjordsgatan 22, 164 40 Kista, Sweden

[fredrik.carlsson|amaru.cuba.gyllensten|magnus.sahlgren]@ri.se

²Gavagai, Kungsgatan 60, 111 22 Stockholm, Sweden

fredrik.olsson@gavagai.io

Abstract

This paper introduces a long-range multiple-choice Question Answering (QA) dataset, based on full-length fiction book texts. The questions are formulated as 10-way multiple-choice questions, where the task is to select the correct character name given a character description, or vice-versa. Each character description is formulated in natural text and often contains information from several sections throughout the book. We provide 20,000 questions created from 10,000 manually annotated descriptions of characters from 177 books containing 152,917 words on average. We address the current discourse regarding dataset bias and leakage by a simple anonymization procedure, which in turn enables interesting probing possibilities. Finally, we show that suitable baseline algorithms perform very poorly on this task, with the book size itself making it non-trivial to attempt a Transformer-based QA solution. This leaves ample room for future improvement, and hints at the need for a completely different type of solution.

1 Introduction

Comprehending and analyzing fictional stories plays an important part in human culture (Smith et al., 2017). In particular, book studies is a commonly applied educational tool used to both probe and enrich students’ language comprehension skills (Tunnell and Jacobs, 1989). Ideally, these book studies require the students to reason about notions spread out over hundreds of pages of text.

By contrast, methods and datasets for machine reading comprehension (MRC) have predominantly been limited to comparably short texts, with evaluations often focusing on various forms of short-text natural language understanding (NLU) tasks, where the input is limited to a small number of sentences. Examples of such tasks include textual similarity (Agirre et al., 2012), sentiment

analysis (Yu and Jiang, 2016), Question Answering (Yadav et al., 2019), inference and, entailment (Talman et al., 2019), etc.

There is an ongoing debate regarding NLU evaluation datasets, spurred by the acclaimed superhuman results on benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). The critique points out several inherent issues with these benchmarks (Tsuchiya, 2018), such as data-leakage (Elangovan et al., 2021), and that models are sometimes able to “cheat” by exploiting spurious linguistic cues in the data (Niven and Kao, 2019). Proposed mitigation methods include the use of adversarially hard datasets (Nie et al., 2020), and taking a more rigorous approach to dataset design (Bowman and Dahl, 2021).

Adding our voice to this discussion, we point out an additional limitation to near all these prior datasets, namely how limited they are in the amount of text that is used per question (See section 2 for notable exceptions). Since datasets arguably drive the direction of research, we find it backward and stagnating to only create evaluation tasks suitable for the current paradigm of fixed-size Transformer models (Vaswani et al., 2017)). Therefore, we find it equally important to create datasets meant to task methods on long-text comprehension.

Thus, we present **GANDALF** (a **General chAracter Name Description dAtaset for Long Fiction**), a full-length book Question Answering dataset focused on the novel task of **character description recognition**. This 10-way multiple-choice task asks the model to read *an entire book*, and then identify the correct character name to a given character description, or vice-versa. In total, we supply 20,000 questions, with a 50/50 split between predicting a name given a description, and predicting the description given a character name. The manually created descriptions are all expressed in natural text and contain a mixture of traits, important events, and relationships to other characters.



		Desc 2 Name	
 <p>“ Chapter 1: This text is about Dante and can be replaced with the real first passages of the Divine Comdey. Hence, you can discard what it currently says. Finally, Dante woke up and realised that everything, had been a bad dream. THE END ”</p>	+	“Acts as Dante’s guide and is symbolic for human reason”	1 Julius Ceasar 2 Virgil . . . 10 Dante Alighieri
		Name 2 Desc	
 <p>“ Chapter 1: This text is about Dante and can be replaced with the real first passages of the Divine Comdey. Hence, you can discard what it currently says. Finally, Dante woke up and realised that everything, had been a bad dream. THE END ”</p>	+	Dante Alighieri	1 “The protagonist” 2 “Dante’s guide in Paradisio” . . . 10 “Works on behalf of Beatrice”

Figure 1: Example of questions formulated as *Desc2Name* and *Name2Desc*.

A schematic illustration of GANDALF is provided in Figure 1.

Taking into account the current discourse concerning datasets, we implement a simple name-replacement system that counters potential data leakage. This system also enables for straightforward implementations of probing tasks by controlling for example gender or nationality. Finally, we perform experiments intended to measure a model’s ability to cheat on GANDALF, by answering the questions without relying on the book data.

The full dataset is available at: github.com/FreddeFrallan/GANDALF-Dataset

2 Related Work

There exists a rich body of literature for various MRC and NLU-related tasks. Examples of previous MRC datasets that also formulate their questions as multiple-choice include: RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), MultiRC (Khashabi et al., 2018) and RACE-C (Liang et al., 2019). For each of these datasets, the combined length of each question and its provided information often fall below 50 sentences, and for some datasets significantly less.

Related work which specifically utilize books as their universe of knowledge include Children’s Book Test (CBT) (Hill et al., 2016), BookTest (Bajgar et al., 2016), and COMICS (Iyyer et al., 2017). These three datasets all utilize cloze-style answers from sequences with up to 21 sentences.

NarrativeQA (s Kořiský et al., 2018) provide questions based on full-length stories, with an average of ~60k words per story. The answer format is free-text and is hence evaluated using text similarity metrics such as BLEU (Papineni et al., 2002) and Meteor (Banerjee and Lavie, 2005).

In contrast to previous work, the books within GANDALF contain on average ~150,000 words and ~8,000 sentences. Each question is classified into its level of referential complexity (See section 3.2), which when combined with the multiple-choice accuracy results in an informative evaluation metric. To the best of our knowledge, GANDALF is therefore not only the longest current MRC dataset, but also the only current MRC dataset which provides these insights during evaluation. We note that there are other types of benchmarks and tasks that requires long context, such as Thorne et al. (2021) and Huang et al. (2021).

2.1 The Dataset Discourse

Recently, the robustness and validity of many academic NLU datasets have come into question. Kaushik and Lipton (2018) proves that many questions in existing benchmarks can be solved without even considering the corresponding context. Paullada et al. (2020) identify multiple shortcomings in the common practices for dataset collection prevalent within the machine learning community. Elangovan et al. (2021) point to a substantial data leakage between train and test data for many datasets.

The data-leakage problem recently surfaced in the context of MRC, when the long-form dataset ELI5 (Fan et al., 2019) received critique for leaking at least 81% (Krishna et al., 2021). To avoid such shortcomings, Kaushik and Lipton conclude that researchers must validate that both the question and their respective context are required for solving the task; and be cautious when using cloze questions.

Finally, Bowman and Dahl (2021) present four criteria that NLU datasets should meet in order to make evaluation reliable. These criteria state that task performance needs to be highly correlated

	Level 1	Level 2	Level 3	Level 4	All
Total number of questions	1,700	3,700	2,300	2,300	10,000
Mean number of sentences	1.05	1.05	1.16	1.17	1.10
Mean number of words	13.53	13.46	28.37	28.47	20.35
Total number of tokens	64,649	6,908	5,298	5,048	9,729

Table 1: Statistics for the 10,000 character descriptions, structured by their level of referential complexity.

Avg number of characters	26.93
Max number of characters	89
Min number of characters	10
Total number of characters	4,766
Avg number of sentences	8,370
Avg number of words	152,917
Total number of tokens	27,066,319

Table 2: Statistics for the 177 full-length books included in GANDALF and its described characters.

with in-domain performance. Datasets need to be harder and/or bigger. Questions need to be unambiguously labeled, and the dataset should reveal potential biases of the system solving the task.

3 The GANDALF Dataset

GANDALF contains 20,000, 10-way multiple-choice questions, formulated from 10,000 manually created character descriptions of 2,500 characters from 177 books. For each question, the relevant book is provided, and the task is to either predict the correct name given a description or predict the correct description given a character name. For brevity we refer to these two different settings as: *Desc2Name* and *Name2Desc*. See Figure 1 for a visual example.

Additionally, GANDALF contains a simple anonymization system, where names are replaced in both the books and the descriptions. This acts a mitigation to potential data leakage and also allows for easy creation of future probing tasks.

3.1 The Books

The dataset comprises a total of 177 full-length books, collected from Project Gutenberg as described in Section 4.1. Table 2 summarizes basic statistics about the books, and Figure 2 shows the length of the books in number of words. There is one book that is significantly longer than the others (*War and Peace* by Leo Tolstoy), and two books that are shorter than all the others (*In Our Time* by Ernest Hemingway, and *The Queen of Spades* by

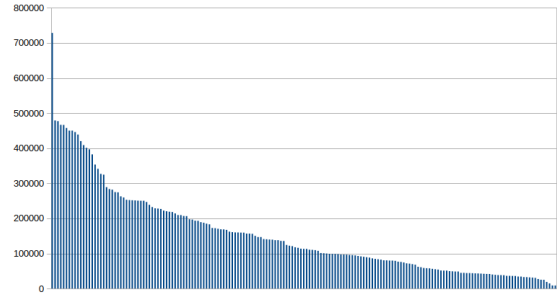


Figure 2: Length of the 177 books in number of words.

Alexander Sergeievith Poushkin). Appendix-E list all book titles within GANDALF.

3.1.1 The Characters

In total, GANDALF includes 4,766 named characters who all match the filtering criteria stated in Section 4.2. For these characters, it was possible to supply uniquely identifiable descriptions for 4,463. The remaining 303 characters who were not given descriptions were however included as potential question alternatives for the *Desc2Name* setting. Table 2 includes some basic statistics about the number of characters included per book.

3.2 The Character Descriptions

In total GANDALF contains 10,000 unique character descriptions of varying complexity. Each description is expressed in a short passage of natural text, spanning 1-2 sentences. These descriptions contain combinations of traits, events, and relationships, which together are meant to uniquely identify a character within its respective book universe. The annotator instructions for the creation of these descriptions are available in appendix C.

These character descriptions are all structured by what we refer to as, its level of "*referential complexity*". The referential complexity is a means to describe the number of deductive steps required to understand a description. The relevant levels of referential complexity are defined in the following list. Examples of different referential complexities are available in Appendix 6, Table 1 displays basic

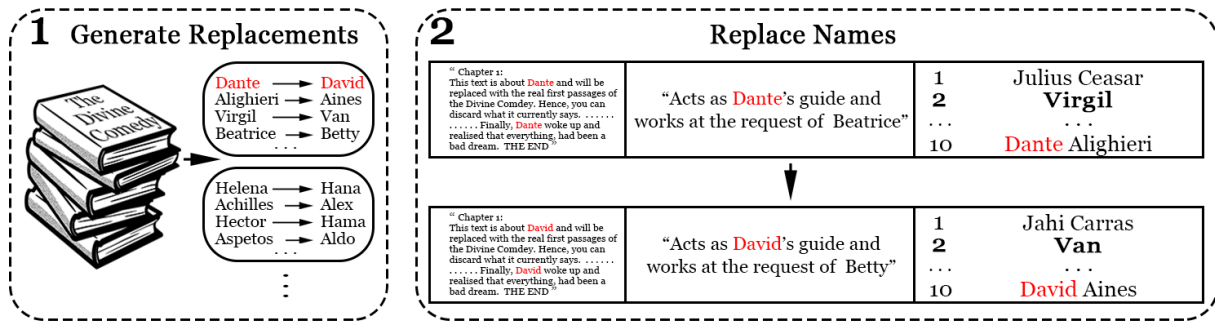


Figure 3: Illustration of the usage of the name-replacement schema included with GANDALF.

statistics for all descriptions, structured by their level of referential complexity.

Level 0: The character is described directly by its own name.

Level 1: The description is self-contained; it contains no references to other characters.

Level 2: The description contains a reference to at least one other character, by stating that character’s name (Level 0).

Level 3: The description contains a reference to at least one other character, by providing a Level 1 description of that character

Level 4: The description contains a reference to at least one other character, by providing a Level 2 description of that character.

We stress that referential complexity does not necessarily correlate with the difficulty of a question. For example, the level 1 description *"The protagonist"*, is expected to be more difficult than the more concrete level 2 description *"The Dog of Dorothy"*. Instead, increasing the referential complexity is a simple way to generate more questions from a fixed set of descriptions which include reference to other characters (See section 4.1).

3.3 Name Replacement Schema

GANDALF incorporate a simple name-replacement schema, illustrated in Figure 3. This schema creates a per book lookup table of name replacements, for all unique *unambiguous name terms* of described characters in that book. Name replacements are assigned either randomly, or according to a heuristic. Similarly to Hutchinson et al. (2012) and Bonato et al. (2016), we find occurrences of all original names by exact matching, and do this for both the descriptions and

books. Finally, all instances of these names are replaced by their newly assigned names. Further details are available in Appendix B.

3.4 Probing Variations

Due to the formulation of the character description recognition task, name replacement enables for straightforward probing variations without having to change the nature of the task. This is attractive since it allows for investigation of how model performance is affected by slightly altering the data while keeping both task formulation and model constant. By contrast, most existing probing tasks and datasets formulate probes as a separate classification or inference task, requiring a model to add an additional classifier specifically for the probe. Suggestions for future probing tasks are available in Section 7.1.

4 Creating the dataset

The creation of the GANDALF dataset is a process that in large is guided by avoiding copyright infringement (See section 7.3). Figure 4 along with the following list gives an overview of the creation process, upon which the following subsections describe each step in more detail.

1. Collect a large set of character essays for books available in the public domain.
2. Discard essays for characters that are not described directly by a name.
3. Manually extract descriptive traits, events, and relationships from the character essays.
4. Manually combine the extracted information into level 1 and level 2 character descriptions.
5. Use the already created descriptions to manually create level 3 and level 4 descriptions.

6. Generation *Name2Desc* and *Desc2Name* questions.

4.1 Collecting the data

At the initial step of creating GANDALF, we collected a large set of character essays from various online sources. To make the dataset free for redistribution, we only gathered essays corresponding to books that are currently available in the public domain. A full list of the sources used for character essays is available in appendix 5. Finally, all relevant books were downloaded from Project Gutenberg via the unofficial API¹.

4.2 Filtering the data

After the collection phase, all data was filtered to fit the following criteria: All characters must have at least one part of their full name suitable for name replacement, and each book must contain at least 10 characters. This entails that we discard all essays which do not explicitly refer to a single entity and whose name does not contain at least one unambiguous name term (See Appendix B). Examples of character essays discarded are thus: *"The narrator"*, *"Doctor Justice"*, and *"The Wilson twins"*.

4.3 Extracting character information

From the remaining character essays, annotators manually extracted sequences of text with descriptive character information. The annotators were instructed to find the perceived gender and text sequences containing at least one of the following: character traits, descriptive book events, or relation to another character.

4.4 Creating level 1 & level 2 descriptions

Using the extracted character information, annotators manually composed short uniquely identifiable level 1 and level 2 descriptions. Taking extra care to formulate descriptions that did not contain any of the respective characters' name terms. To determine the information required to identify a character, annotators cross-compared described characters within the same book. The annotators thus worked exclusively with the character essays.

4.5 Creating level 3 & level 4 descriptions

Level 3 and level 4 descriptions were manually created from the already created level 1 and level 2 descriptions. reformulating existing character

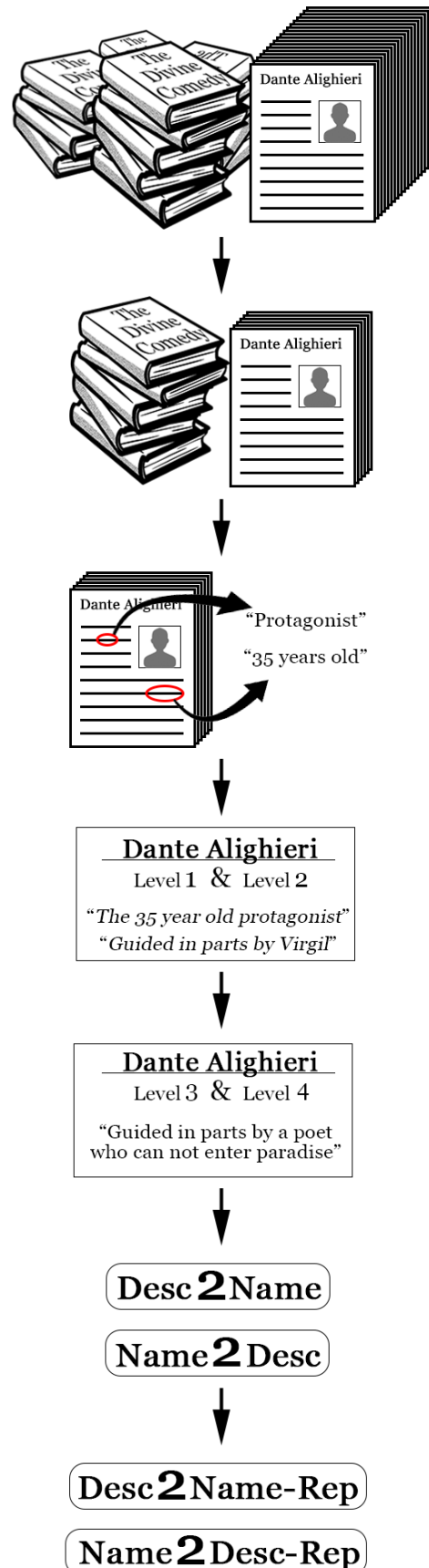


Figure 4: Illustration of the overall creation process of GANDALF.

¹<https://github.com/c-w/gutenberg>

references in accordance with the definitions in section 3.2. This resulted in an additional 4,600 descriptions of 2,356 characters.

4.6 Generating Questions

The 10,000 character descriptions were used to generate two question sets: *Name2Desc* and *Desc2Name*. For *Desc2Name* the incorrect question alternatives were created by randomly selecting names or descriptions for other characters from the same book. This results in 10,000 *Name2Desc* questions and 10,000 *Desc2Name* questions.

4.7 Name Replacement

Finally, we include name-replaced versions of the two basic settings of the dataset, named *Name2Desc-Rep* and *Desc2Name-Rep*. These are created by shuffling the names among all characters from all books, controlling for first and last names. The name-replacement tables are hence generated by mapping each first name to another first name of a character from any book within GANDALF, and the same for last names.

5 Experiments

To the best of our knowledge, there has been no proposed method that is capable of handling naturally formulated questions along with the long-range format of GANDALF. We are therefore limited in what experiments we can perform without proposing a new solution, which is deliberately not within the scope of this paper. Therefore, our experiments can be interpreted more as testing the robustness of our dataset, while they also demonstrate that the problem is non-trivial.

All three of our approaches are based on selecting the alternative which maximizes the probability of the description being stated after the character name. The first method is a traditional word-based statistical method, and the other two utilize modern language models. The statistical model acts as a simple information-retrieval baseline, and a way to measure potential noise that could be introduced during name replacement. The language model approaches are intended as an attempt to utilize off-the-shelf NLU technology to solve the problem both the intended way, and without the book texts.

Finally, we saw no increase in performance after fine-tuning the language models, the reported results are hence attained from directly applying

pre-trained checkpoints.² These experiments therefore include all the questions of GANDALF as test data.

5.1 BoW: TF-IDF

As a simple information-retrieval baseline we perform a TF-IDF (Salton and McGill, 1986) search over all paragraphs in a target book, for all description-queries. This entails that we gather TF-IDF-statistics for each book text and description-query, constructing sparse TF-IDF representations for all book paragraphs, as for all queries. Finally, we compare the cosine similarity between each description-query and each paragraph, and select the query with the highest similarity to any paragraph.

5.2 Causal Language Model: Transformer-XL

Transformer-XL (Dai et al., 2019) is one of the few transformer-based language models without a fixed input size, making it a viable contender for GANDALF. We are hence able to state the full book text as context, prior to posing the different queries. This is achieved by first computing the residual memory for the complete book text, and then providing that alongside every description-query.

5.3 Causal Language Model: GPT-2

GPT-2 (Radford et al., 2019) has a fixed input size of 1,024 tokens, making it unable to comprehend the full book texts. However, it has been trained on a vastly large dataset of text scraped from the internet. This makes it a suitable model for measuring potential data leakage and other potential lexical artifacts which might make the questions trivial by themselves. Especially, since it is highly likely that GPT-2 has been trained on both the books and the original character essays which are used to create GANDALF.

6 Results

Table 3 displays the accuracy of the three different baselines, for the 4 different standard versions of GANDALF. Although, neither of the methods produces any good results and lie very close to the random baseline, the TF-IDF approach performed the best. Hence, table 4 is included, that show

²Checkpoints are taken from https://huggingface.co/transformers/pretrained_models.html

	Name2Desc	Desc2Name	Name2Desc-Rep	Desc2Name-Rep
Random	10.0	10.0	10.0	10.0
Book + Query				
Transformer-XL	12.9	9.6	12.4	10.1
BoW + TF-IDF	19.9	14.4	19.8	13.9
Query Only				
GPT-2	9.4	11.5	9.8	11.7
GPT-2 Medium	9.5	12.2	9.4	11.8
GPT-2 Large	10.5	12.4	10.5	12.0
GPT-2 XL	10.5	11.6	10.8	12.0

Table 3: Model accuracy on the four different versions of GANDALF.

	Level 1	Level 2	Level 3	Level 4
BoW + TF-IDF				
Name2Desc	19.9	22.9	4.6	6.3
Desc2Name	21.6	20.9	17.4	19.7
Name2Desc-Rep 2	19.2	22.4	4.6	5.7
Desc2Name-Rep	20.3	20.7	17.2	19.5

Table 4: BoW + TF-IDF accuracy on the different levels, on all four different versions of GANDALF.

the per-level accuracy over the different referential complexities for TF-IDF.

6.1 Book + Query

Transformer-XL performs nearly on par with random, although there is a very slight improvement on both *Name2Desc* tasks, compared to both random and the query only approach. This lack of performance demonstrates that current long-range Transformers struggle with the book texts of GANDALF.

The TF-IDF-based approach displays a notable performance increase, achieving the best results in all settings. A notable difference of 5.5 and 5.9 points can be seen between *Name2Desc* and the *Desc2Name* counterpart. The results in table 4 clearly show that this difference is due to TF-IDF being incapable of handling level 3 and level 4 questions in the *Desc2Name* setting.

Finally, the difference between the normal and the name-replaced datasets, are for both methods near negligible. We stress that this is the *desired* result, as this indicates that most statistical properties remain intact through the alteration of name replacement. Hence allowing for the deployment of various renaming schema, as discussed in section 7.

6.2 Query Only

Turning to GPT-2, which discards all book texts, performance are again very close to the random baseline. Both *Desc2Name* sets do however see an increase of circa 2 points compared to *Name2Desc*, and results tend to increase by circa 1 percentage point going from smaller to larger models on all sets.

This small difference might be negligible, but it could also indicate that the *Name2Desc* setting is more prone to lexical artifacts, which makes an inductive guess better than random.

7 Future Work & Discussion

Systems specialized towards a single task attain poor generalization abilities, and hence demonstrate low levels of intelligence (Chollet, 2019). As AI researchers, our main interest is therefore not a method capable of **only** solving the character recognition task of GANDALF. Rather, our ultimate goal is methods capable of handling and generalizing over a wide range of tasks, domains, and modalities.

Current models perform well over the different tasks included in benchmarks such as GLUE and SuperGLUE, but they are yet capable of handling both long and short texts. Therefore, we think the time is right for extending our current evaluation benchmarks to include tasks covering large bodies

of text. GANDALF with its potential extensions and probing tasks is hence our first contribution to such a set setup.

We note the relatively weak performance of the models tested in our experiments. It is probable that better performance can be attained using alternative techniques, such as the Dense Passage Retriever (Karpukhin et al., 2020) or Retrieval Augmented Generation (Lewis et al., 2020). We leave this for future work.

7.1 Extensions to GANDALF

Two straightforward types of probing variations that the GANDALF data enables is to study a model’s sensitivity to gender and racial bias. In the case of gender bias, we can simply switch the gender of all names and study how this affects the performance. Another possibility is to replace *all* character names with male or female names. In the case of racial bias, we can replace character names with typical names of some specific demographic group.

It is also straightforward to include negated statements in the character description, enabling studies of models’ sensitivity to negation (Ettinger, 2020). Such negated statements can be produced by simply selecting descriptions from other characters, possibly in the same book, and negating them (“*is not* the dog of Dorothy” or “*does not* act as Dante’s guide”).

7.2 Character Description Recognition

Although we are not personally interested in methods that only aim to solve GANDALF’s character description recognition task, we recognize that others might be. We advise researchers wishing to pursue such solutions, to combine existing NLP methods utilized in data-driven literature studies. For example, extracting character networks (Labatut and Bost, 2019) would intuitively be useful for solving questions involving character relations. Additionally, certain rule-based heuristic might also prove useful, as it is likely that a character labeled as “*The Protagonist*” will have the highest frequency of occurrences throughout the book. Finally, we note that the work of (Zhang et al., 2019) focus specifically on automatically generating character descriptions from books (Unfortunately published without accompanying code).

7.3 Copyright Protection

To ensure that legalities do not interfere with scientific reproducibility, we stress the importance of having a freely distributable dataset. GANDALF only includes books that are available in the public domain, and facts are not covered by copyright protection. So while the original character essays themselves might be under copyright protection, the facts they express are not. Hence, both our collected set of books and our generated set of questions are free to be publicly distributed for future research.

8 Conclusion

This paper has introduced the GANDALF dataset, which constitutes a unique challenge for machine reading comprehension by requiring long-range retention capabilities while simultaneously being able to ignore irrelevant information. We have introduced the character description task with its two variations (Desc2Name and Name2Desc), and argued that this task formulation provides unique opportunities for probing variations without changing the nature of the task itself. We also provide a number of baseline results on the dataset using both simple and more advanced methods, and the results clearly demonstrate the challenging nature of the dataset.

We believe that this dataset and task provides a welcome addition to existing machine reading comprehension benchmarks, in particular at a time when we start to see superhuman performance on existing datasets, with an apparent risk of models starting to optimize for a specific dataset rather than for general reading comprehension abilities. The GANDALF dataset, by contrast, is extremely challenging with minimal risk of data leakage and consequently low risk of models cheating on the tasks.

References

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. [SemEval-2012 task 6: A pilot on semantic textual similarity](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

- Ondrej Bajgar, Rudolf Kadlec, and Jan Klein-dienst. 2016. [Embracing data abundance: BookTest Dataset for Reading Comprehension](#). *arXiv:1610.00956 [cs]*. ArXiv: 1610.00956.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Anthony Bonato, David Ryan D’Angelo, Ethan R. Elenberg, David F. Gleich, and Yangyang Hou. 2016. Mining and modeling character networks. In *Algorithms and Models for the Web Graph*, pages 100–114, Cham. Springer International Publishing.
- Samuel R. Bowman and George E. Dahl. 2021. [What Will it Take to Fix Benchmarking in Natural Language Understanding?](#) *arXiv:2104.02145 [cs]*. ArXiv: 2104.02145.
- François Chollet. 2019. [On the measure of intelligence](#).
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. [Transformer-XL: Attentive language models beyond a fixed-length context](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.
- Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. [Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance Evaluation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.
- Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long Form Question Answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. [The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations](#). *arXiv:1511.02301 [cs]*. ArXiv: 1511.02301.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. [Efficient attentions for long document summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.
- Sterling Hutchinson, Vivek Datla, and M. Louwerse. 2012. Social networks are encoded in language. *Cognitive Science*, 34.
- Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé, and Larry Davis. 2017. [The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487. ISSN: 1063-6919.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Divyansh Kaushik and Zachary Chase Lipton. 2018. [How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks](#). *EMNLP*.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. [Hurdles to Progress in Long-form Question Answering](#). *arXiv:2103.06332 [cs]*. ArXiv: 2103.06332.
- Vincent Labatut and Xavier Bost. 2019. [Extraction and analysis of fictional character networks](#). *ACM Computing Surveys*, 52(5):1–40.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020.

- Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 742–757, Nagoya, Japan. PMLR.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *arXiv:2012.05345 [cs]*. ArXiv: 2012.05345.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Tomáš Kořický, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, TBD:TBD.
- Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.
- Daniel Smith, Philip Schlaepfer, Katie Major, Mark Dyble, Abigail Page, James Thompson, Nikhil Chaudhary, Gul Deniz Salali, Ruth Mace, Leonora Astete, Marilyn Ngales, Lucio Vinicius, and Andrea Migliano. 2017. Cooperation and the evolution of hunter-gatherer storytelling. *Nature Communications*, 8.
- Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. 2019. Sentence embeddings in nli with iterative refinement encoders. *Natural Language Engineering*, 25:467–482.
- James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. Database reasoning over text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3091–3104, Online. Association for Computational Linguistics.
- Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).
- Michael O. Tunnell and James S. Jacobs. 1989. Using "real" books: Research findings on literature based reading instruction. *The Reading Teacher*, 42(7):470–477.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3266–3280. Curran Associates, Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Alignment over heterogeneous embeddings for question answering. 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 2681–2691, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas. Association for Computational Linguistics.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. [Generating character descriptions for automatic summarization of fiction](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7476–7483.

```
https://www.cliffsnotes.com/  
https://www.coursehero.com/  
https://www.gradesaver.com/  
https://www.sparknotes.com/
```

Table 5: All the sources used for gathering the character essays used to create GANDALF.

A Sources for Character Essays

Table 5 contains the four sites from which character essays were collected.

B Replacement of names

To mitigate the amount of noise introduced to the book texts during name replacement, we aim to only replace terms within character names which unambiguously refer to a name. A character referred to as "*Doctor Emily Bender*", would therefore include the unambiguous name terms "*Emily*" and "*Bender*".

For all GANDALF characters annotators selected their unambiguous name terms, and classified them as first or last names. This was achieved by a combination of manual inspection and querying of WordNet (Fellbaum, 1998) for each name term. Ultimately annotators were allowed to overrule the fact that Wordnet deemed a word ambiguous, if the annotator suspected the other word meanings to be highly unlikely to occur within the respective book. For example the word "*Teta*" also have the following Wordnet definition: "*a member of the large western branch of Sioux people which was made up of several groups that lived on the plains*"

For the finding of name occurrences, we used direct string matching against if a name term occurred as an isolated word, or in combination with a suffix such as *'s*. Admittedly, this does not handle a lot of the many potential corner cases. For example it does not handle the initials of a name, if the name is spelled differently during a stuttering conversations, or if the original name takes part in a word-pun. These name replacements therefore contribute to a certain level of noise to the data.

However, it is our belief that these corner cases formulate the exception rather than the rule. Even when a character is being referred to by a nickname the majority of the time, human readers easily connect the two names to the same entity. Intuitively, we therefore believe that a theoretically super intelligent system, could on many occasions be able to figure out what name replacements went wrong.

C Annotator Instructions

The annotators were tasked to work on a per-book basis, and work from the assumption that the collected character essays contained all essential information required to make a distinguishable character description. This assumption does not necessarily always hold, but it relieves the annotators from having to read the actual book.

First the annotators were asked to extract character traits and descriptions from the collected character essays of a single book. After this information had been extracted, they were then tasked to puzzle together the extracted traits into short descriptions, which had to *uniquely* identify the characters within selected character alternatives from that book. The annotators were told to discard any character that ended up with ambiguous descriptions.

D Character Descriptions

Table 6 contains examples of character descriptions of for different referential complexity levels.

E Books

Table 7 lists all 177 book titles contained within the GANDALF dataset.

Ref Complexity	Book Title	Character Name	Description
1	A tale of Two cities	Roger Cly	An English spy who fakes his death.
1	A tale of Two cities	Sydney Carton	Has fallen into a life of alcoholism.
1	Oliver Twist	Oliver Twist	Is the protagonist of the story.
1	Oliver Twist	Mrs. Maylie	A lady, who is very dignified and stately, despite her old age.
1	Moby Dick	Starbuck	Thinly built Quaker who displays a pragmatic manner.
1	Moby Dick	Stubb	The second mate of the Pequod.
2	Silas Marner	Eppie	The catalyst who integrates Silas into Raveloe.
2	Silas Marner	Godfrey Cass	Secretly married to Molly Farren.
2	Crime and punishment	Katerina	Consumptive wife of Marmeladov.
2	Crime and punishment	Andrei Lebezyatnikov	Grudging roommate of Luzhin.
2	The brother's Karamazov	Dmitri Karamazov	Eldesst son of Fyodor by his first wife.
2	The brother's Karamazov	Mussyalovich	A Polish officer who betrays Grushenka.
3	Call of the wild	Spitz	Experienced, clever, and fears and hates the protagonist.
3	Call of the wild	Hans	A partner of the person who does not fear the wild because he knows how to survive it.
3	My Antonia	Jan Cuzak	The shy son of the person who stays in the country, remarries and has many children.
3	My Antonia	Joe	Patient with his mother-in-law.
3	The Dubliners	Mary Jane	Cousin to the protagonist in "The Dead".
3	The Dubliners	Gallaher	Friend to the character who daydreams while working a desk job in "A Little Cloud".
4	Winesburg, Ohio	Mook	The only person who the son of Ebenezer is comfortable talking to.
4	Winesburg, Ohio	Katherine Bentley	David Hardy's grandfather's wife.
4	The Tenant of Wildfell hall	Mr Leighton	Preacher at the church where Arthur's wife attends.
4	The Tenant of Wildfell hall	Mrs. Wilson	Mother to the person who spreads rumors about Helen.
4	Martin Chuzzlewit	Ruth Pinch	Sister of the very likable, ex-student who works for Mr Pecksniff.
4	Martin Chuzzlewit	Mark Tapley	Friend and Valet of the stubborn architectural student of Mr Pecksniff.

Table 6: Example descriptions of varying referential complexity.

<p>A connecticut yankee in king arthur's court A portrait of the artist as a young man A room with a view A vindication of the rights of woman An essay on the principle of population Anna karenina Barnaby rudge Charlotte temple Cranford David copperfield Dialogues concerning natural religion Dracula Emma For the term of his natural life Gilgamesh He knew he was right Henry vi, part 1 Henry viii In our time Kidnapped Lady susan Little dorrit Madame bovary Major barbara Mary barton Moby dick My antonia O pioneers! Orlando Pamela, or virtue rewarded Richard ii Roughing it Silas marner Summer The adventures of sherlock holmes The age of innocence The autobiography of benjamin franklin The brothers karamazov The castle of otranto The flowers of evil The history of the peloponnesian war The hound of the baskervilles The hunchback of notre dame The jungle The man of the forest The moonstone The odyssey The private memoirs of a justified sinner The queen of spades The school for scandal The souls of black folk The tenant of wildfell hall The vicar of wakefield This side of paradise Titus andronicus Troilus and criseyde Ulysses Villette Winesburg, ohio</p>	<p>A hero of our time A midsummer night's dream A study in scarlet Adam bede American notes Anne of green gables Black beauty Common sense Crime and punishment Demian Dombey and son Dubliners Ethan frome Frankenstein Gulliver's travels Henry iv, part 1 Henry vi, part 2 History of tom jones, a foundling Jane eyre King henry iv, part 2 Lady windermere's fan Little women Maggie: a girl of the streets Mansfield park Meditations Mrs dalloway An American Slave Oliver twist Othello Peter pan Richard iii Second treatise of government Sister carrie Swann's way The adventures of tom sawyer The ambassadors The awakening The call of the wild The country of the pointed firs The frogs The golden asse The house of mirth The idiot The jungle book The mayor of casterbridge The mysteries of udolpho The orkneyinga saga The portrait of a lady — volume 1 The republic The secret agent The spanish tragedy The trial The war of the worlds Three men in a boat To the lighthouse Twelfth night Uncle tom's cabin War and peace Women in love</p>	<p>A journal of the plague year A passage to india A tale of two cities Adventures of huckleberry finn An ideal husband Antony and cleopatra Candide Coriolanus Daniel deronda Desert gold Don juan Edgar huntly Far from the madding crowd Germinal Hard times Henry v Henry vi, part 3 Howards end Jude the obscure Lady audley's secret Little brother Lord jim Main street Martin chuzzlewit Middlemarch Much ado about nothing North and south On the origin of species Our mutual friend Ragged dick Romeo and juliet Siddhartha Sons and lovers Tarzan of the apes The aeneid The american The beautiful and damned The canterville ghost The duchess of malfi The gilded age The good soldier The house of the seven gables The jew of malta The king in yellow The mill on the floss The mysterious affair at styles The pickwick papers The praise of folly The scarlet pimpernel The sorrows of young werther The subjection of women The valley of fear The woman in white Timon of athens Troilus and cressida Typee Vanity fair White fang Wuthering heights</p>
---	---	--

Table 7: All the 177 book titles contained within GANDALF.