🥞TEXT-CAKE: Challenging Language Models on Local Text Coherence

Luca Dini[1,2], Dominique Brunato[1], Felice Dell'Orletta[1], Tommaso Caselli[3]
[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa
[2]University of Pisa [3]Center for Language and Cognition (CLCG), University of Groningen
{luca.dini, dominique.brunato, felice.dellorletta}@ilc.cnr.it, t.caselli@rug.nl

Abstract

We present a deep investigation of encoder-based Language Models (LMs) on their abilities to detect text coherence across four languages and four text genres using a new evaluation benchmark, 🥞**TEXT-CAKE**. We analyze both multilingual and monolingual LMs with varying architectures and parameters in different finetuning settings. Our findings demonstrate that identifying subtle perturbations that disrupt local coherence is still a challenging task. Furthermore, our results underline the importance of using diverse text genres during pre-training and of an optimal pre-traning objective and large vocabulary size. When controlling for other parameters, deep LMs (i.e., higher number of layers) have an advantage over shallow ones, even when the total number of parameters is smaller.

1 Introduction & Motivations

A text is a semantic unit made of meanings conveying knowledge (Halliday and Hasan, 1976). Under this perspective, text coherence is "the mutual access and relevance within a configuration of concepts and relations" (De Beaugrande and Dressler, 1981). The "mutual access" and the "configuration of relations", however, identify features that relate more to the perception of a text rather than a text itself (Wang and Guo, 2014). This vision highlights the nature of text as a process, where coherence emerges from the interaction between the producer and the receiver requiring mutual effort. While Linguistics considers coherence as one of the "standards of textuality" (De Beaugrande and Dressler, 1981, 3), coherence is inherently a psychological construct.

Given the role of coherence as an essential component of text as well as its use in application-oriented scenarios, from healthcare (Parola et al., 2023; Elvevåg et al., 2007; Iter et al., 2018) to automatic essay scoring in language learning (Lai and Tetreault, 2018; Mesgar and Strube, 2018) and readability assessment (Pitler and Nenkova, 2008; Feng et al., 2009), modeling (text) coherence is an essential task. The rise of pre-trained LMs has revitalized this area of research, offering new avenues for exploration with the emergence of two key trends. The first involves adapting established coherence modeling frameworks by leveraging deep neural networks to enhance performance on various downstream tasks (Muangkammuen et al., 2020; Tien Nguyen and Joty, 2017). The second focuses on interpretability, investigating how neural representations capture discourse phenomena beyond sentence level. As detailed in § 2, these interpretability studies share commonalities in evaluation, often using tasks that act as proxies for coherence assessment, such as recognizing original versus sentence-shuffled texts or detecting an intruder sentence (Chen et al., 2019; Shen et al., 2021).

Our work builds upon existing research by investigating LMs' ability to distinguish coherent from artificially created incoherent text. To this end, we introduce a novel and language-independent method to gradually disrupt the local coherence within a text and evaluate the performance of transformer-based encoders on this task through a set of extensive experiments across different text genres and languages. Specifically, our study makes the following main contributions:[1]

- we present 🥞**TEXT-CAKE**, a new multilingual and multigenre dataset specifically conceived for the task of local discourse coherence detection in written text;
- we conduct an extensive investigation of multiple enconder-based models to assess their ability to identify coherence at various levels

[1] Data and code https://github.com/lucadinidue/coherence_text_cake

Proceedings of the 31st International Conference on Computational Linguistics, pages 4384–4398
January 19–24, 2025. ©2025 Association for Computational Linguistics

of granularity and their sensitivity to different types of textual perturbations;

- we evaluate how models' architectures and parameters impact their performance when finetuned on this task to gain better insights on which combinations work best as well as to identify directions for further development.

The remainder of this paper is structured as follows: in § 2 we discuss previous work, highlighting differences with our approach. In § 3 we introduce our dataset and approach for its creation. Experiments design, results on finetuned LMs are discussed in § 4 and § 5 respectively. Conclusions and recommendations are in § 6.

2 Related Work

LMs have shown impressive abilities when it comes to functional linguistic skills (Mahowald et al., 2024). Yet, many challenges are still pending when it comes to discourse phenomena beyond the sentence level, such as detecting the coherence of a text. A task which is even more relevant in light of the advancements of generative models (Laban et al., 2021).

A requirement for such evaluations is the development of benchmarks designed to target various nuances of discourse related to coherence. Existing benchmarks range from more linguistically-informed tasks, such as labeling the type of discourse relations or determining the most appropriate discourse markers (Koto et al., 2021; Godunova and Voloshina, 2024; Pandia et al., 2021; Nie et al., 2019; Farag et al., 2020), to tasks that ask models to classify a text as coherent or not, or identifying the original order of sentences within a shuffled text (Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Laban et al., 2021; Shen et al., 2021; Koto et al., 2021). Our work is more in line with this latter set of contributions, although we introduce major differences to address some limitations. First, we combine in a single benchmark two coherence tests: the Shuffle Test (Barzilay and Lapata, 2008; Laban et al., 2021) and the Insertion Test (Elsner and Charniak, 2011); second, we challenge models to identify the specific perturbation class against the original text, thus making this a multi-class classification task; third, we focus on local coherence, i.e., coherence in passages of four sentences.

Beyond the formulation of specific tasks, several perspectives have been explored in the literature, aiming at understanding generalization abilities of LMs, the impact of pre-training objectives and models' size, and which layers best capture discourse information. Shen et al. (2021) tested LMs to recognize an intruder sentence in in-domain and out-domain setting. Their results indicate limited generalization capabilities when applied out-of-domain. Focusing on cross-lingual generalization, Kurfalı and Östling (2021) evaluated sentence encoders on a range of discourse-level tasks. Their findings revealed a performance gap between evaluation in the training language and zero-shot evaluation in unseen languages. In contrast, Godunova and Voloshina (2024) did not report a significant drop between high- and low-resourced languages, suggesting that pre-trained encoders may capture some language-independent aspects of discourse structure. A distinguishing feature of our contribution is that we address all these aspects in a comprehensive analysis.

3 🥞TEXT-CAKE: Data & Task Setting

To robustly investigate local coherence in written text, we have developed a new multilingual and multigenre benchmark. We selected four Indo-European languages belonging to two subgroups (English and Dutch for Germanic; Spanish and Italian for Romance). We chose these languages because of i.) the availability of public repositories for all genres; ii.) the presence of these languages in the pre-training data of selected multilingual LMs; and iii.) the presence of comparable sets of grammatical (e.g., anaphoric reference, conjunction, verb tense, among others) and lexical (e.g., synonymy, repetition, collocation) cohesive devices marking the expression of coherence (Halliday and Hasan, 1976; Tanskanen, 2006).

We selected four text genres ranging from highly controlled text types (news articles and Wikipedia pages) to speech transcripts (TED talks) and fictional stories (fanction). These genres show significant variations in style, writing conventions, and consequently, ways in which coherence is expressed and realized. Furthermore, two of them (Wikipedia and news articles) are common pre-training data for LMs. Conversely, according to the official documentation of available encoder-based LMs, none of them has been directly exposed to fanfictions or TED Talks.

Data sources For Wikipedia and TED Talks, we used existing data repositories. In particular, for Wikipedia, we used Wikimedia dumps from September 20, 2023.[2] The transcriptions of TED Talks were sourced from the TED2020 dataset (Reimers and Gurevych, 2020). For the news articles, we resorted to language specific resources. For English, we used the New York Times portion of the English Gigaword Corpus v5.0 (Graff et al., 2003); for Dutch, the DpgMedia2019 corpus (Yeh et al., 2019)[3]; for Spanish we have been obtained data from the Spanish portion of the DACSA corpus (Segarra Soriano et al., 2022) and for Italian, we used the news corpus from Mattei et al. (2020).

We have extracted fanfiction stories from the GOLEM project collection.[4] We selected only fanfiction whose audience is indicated as "General Audience" and, among them, we have chosen those with only one chapter. This choice has been dictated by technical reasons. Since in the GOLEM collection there is currently a lack of a standardized way of indicating the change from one chapter to the other, we would risk considering as coherent text passages of sentences from different chapters.

Task Settings & Data Split Contrary to other tasks (Mostafazadeh et al., 2016; Chambers and Jurafsky, 2009; Granroth-Wilding and Clark, 2016), our goal is not to predict "what comes next" (being it the logical end of a short story or an event), but to investigate the abilities of encoder-based LMs to detect the internal coherence of a text passage. For each document in our data collection, we thus extracted blocks of four consecutive sentences, corresponding to *minimal coherence passages* (Brunato et al., 2023). For Wikipedia and fanfictions, where paragraph boundaries are explicitly marked by blank lines, we further ensured that the sentences composing the passages all belong to the same paragraph. We further avoided the same sentence to appear in multiple passages.

The creation process of the dataset has been driven by the way we formulated the task of coherence detection. We have thus implemented two different perturbations. The first, substitution (**Sub**), breaks the internal coherence of a passage by substituting one of the sentences with another one *from the same document* but further away. On the basis of a preliminary analysis of the dataset, we have quantified this distance in the 10th following sentence with respect to the one to be substituted. This perturbation can be applied to each of the four sentences, originating four perturbation classes: *Sub_1*, *Sub_2*, *Sub_3* and *Sub_4*, where the number indicates the substituted sentence.

Dataset	Split	IT	ES	NL	EN
Wikipedia	Train	51.7%	64.0%	53.4%	66.7%
	Test	53.4%	68.1%	28.7%	67.7%
News	Train	31.6%	17.5%	14.8%	16.2%
	Test	30.5%	16.5%	11.7%	16.5%
Ted Talks	Train	3.6%	3.8%	3.4%	4.1%
	Test	3.5%	3.8%	3.1%	4.4%
Fanfiction	Train	–	–	–	56.1%
	Test	–	–	–	54.8%

Table 1: Percentage of passages in train and test starting with the first sentence of a paragraph. Lacking an internal division into paragraphs, for TED Talks and News it indicates the percentage of paragraphs beginning with the first sentence of the document. IT = Italian; ES = Spanish; NL = Dutch; EN = English.

The second perturbation, swapping (**Swap**), focuses on the internal structure of the text passages. Similarly to the Shuffle Test (Barzilay and Lapata, 2008), we change the order of two sentences in the passage. In our case, this results in six possible swapping perturbations: *Swap_1_2*, *Swap_1_3*, *Swap_1_4*, *Swap_2_3*, *Swap_2_4* and *Swap_3_4*, where the numerical indices indicate which sentences are swapped.

Considering our perturbations, we have automatically generated multiple train and test splits composed by four sentence passages for each of the text genres and language in analysis. For all languages, we obtained data for Wikipedia, the news articles, and the TED Talks. For the fanfictions, we obtained data comparable in size only for English. The number of extracted passages per language per genre, before applying the perturbations, is 10,000.

Within this framework, we created two distinct datasets: A coarse-grained dataset with three labels —*Orig* (for unperturbed text passages), *Swap*, and *Sub*. This dataset does not differentiate the positions of the perturbed sentences, and is designed to evaluate model performance in a broader three-class classification task. A fine-grained dataset, which distinguishes between the 4 types of *Sub* perturbations and the 6 types of *Swap* perturbations. In the coarse-grained dataset, the specific type of *Swap* or *Sub* perturbation is randomly selected

[2] https://dumps.wikimedia.org/backup-index.html

[3] https://github.com/dpgmedia/partisan-news2019

[4] https://golemlab.eu/about/

from a uniform distribution. For each language and genre, we generated 30,000 passages for each language and genre, with 24,000 for training and 6,000 for testing, resulting in a total of 370,000 text passages. For the fine-grained dataset, we applied all possible perturbations to each of the 10,000 text passages, producing a total of 1,430,000 text passages. This corresponds to 110,000 passages per language and genre, with 88,000 passages allocated for training and 22,000 for testing. We ensured no contamination between the training and test data.

🥞**TEXT-CAKE** is currently the largest benchmark for local text coherence. Table 1 reports the number of passages starting with the first sentence of a document to address potential biases. Notably, only the Wikipedia and Fanfiction portions exhibit a majority of passages with this property, an important consideration when analyzing the results.

4 Experimental Settings

We have conducted two blocks of experiments. First, we investigated multiple multi-lingual models with the same basic architecture but different vocabulary size, pre-training objectives, total number of parameters, and pre-training data (size and sources), using both versions of the datasets. In the second set of experiments, we investigate which other parameters - such as model's depth (i.e., number of layers) and width (i.e., number of attention heads) - impact on the model's performance by keeping fixed the pre-training objective, pre-training data, and size of the vocabulary. For this, we used the fine-grained dataset and a set of monolingual LMs for English presenting high variability of models' architectures, namely the `DeBERTa` family. The choice English only has been dictated by a lack of variability of models in the other languages.

Models Table B in Appendix C summarizes the parameters of all the selected models.[5] The three multilingual models (`mBERT`, `XML-RoBERTa`, and `mDeBERTa-v3`) share the same architecture and underwent training on all languages considered in our study. `mBERT` and `XML-RoBERTa` were both trained using the Masked Language Modeling (MLM) task, with `mBERT` also including the Next Sentence Prediction (NSP) task. In contrast, `mDeBERTa-v3` uses the Replaced Token Detection task. As for their vocabulary, `mDeBERTa-v3` is eight times larger than those of `XML-RoBERTa` and `mBERT`. Lastly, `mDeBERTa-v3` is the smallest LM (86M), followed by `mBERT` (110M), and `XML-RoBERTa` being the largest (250M).

In the monolingual experiments we used all versions of the English `DeBERTa-V3` model, i.e., `DeBERTa-V3-XSmall`, `DeBERTa-V3-Small`, `DeBERTa-V3-Base`, and `DeBERTa-V3-Large`. Although we can expect a better performance of larger models (deeper and wider) (Kaplan et al., 2020; Raffel et al., 2020), He et al. (2023) present an interesting counterpoint for `DeBERTa-V3-XSmall`, showing better results on the MNLI and SQuAD v2.0 benchmarks than `DeBERTa-V3-Small`, despite having only half the parameters (i.e., being narrower). The authors attribute this to `DeBERTa-V3-XSmall`'s deeper architecture, suggesting that a greater number of layers might be crucial for encoding semantics. We want to investigate the validity of this finding for coherence.

Coherence Detection as Classification We frame coherence detection as a multi-class classification task using finetuned models. Following our perturbation settings, each model is challenged to identify which of the classes in 🥞**TEXT-CAKE** is correct given a specific text passage. Given the perfectly balanced nature of our benchmark, models are evaluated using Accuracy.

Before running the main experiments, we conducted trials to determine the optimal settings for finetuning through cross-validation on the training datasets. Details are reported in Table C in Appendix D. All models have been finetuned using the `transformers` library from Hugging Face.[6] The model input is constructed by concatenating the sentences of the text passage, either original or perturbed, and then fit the entire text passage to the model. Class decision is done by adding a classification head on top of the finetuned LMs. As baseline, we have used a Linear SVM classifier trained with uni- and bi-grams. Each passage is represented as the union of the count of its n-grams, with sentences kept distinct to preserve information on sentence order. Random performance corresponds to an Accuracy of 0.33 for the coarse-grained dataset and 0.09 for the fine-grained one.

5 Results

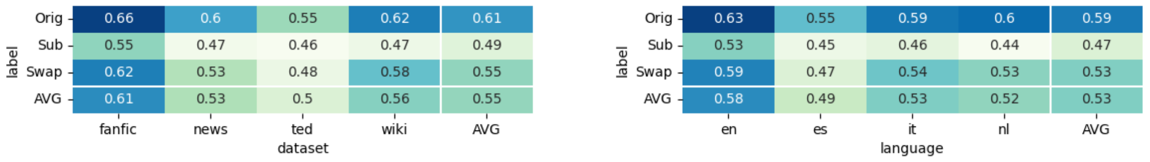
First, we will examine the experiments conducted with the multilingual models (§ 5.1) and subsequently for the monolingual ones (§ 5.2).

[5] All values have been obtained from the models' official documentation.

[6] We used version 4.37.1

		Models			
Dataset	**Language**	`mBERT`	`XML-RoBERTa`	`mDeBERTa-v3`	`Baseline`
Wikipedia	IT	0.59	0.51	**0.62**	0.35
	ES	0.53	0.44	**0.57**	0.34
	NL	0.58	0.53	**0.63**	0.34
	EN	0.57	0.50	**0.60**	0.35
News	IT	0.55	0.55	**0.61**	0.35
	ES	0.40	0.47	**0.56**	0.34
	NL	0.49	0.51	**0.58**	0.33
	EN	0.54	0.56	**0.61**	0.34
TED Talks	IT	0.42	0.48	**0.52**	0.33
	ES	0.44	0.48	**0.55**	0.33
	NL	0.46	0.47	**0.54**	0.33
	EN	0.46	0.50	**0.55**	0.34

Table 2: Results of the multilingual LMs across datasets and languages on the coarse-grained datasets. `Baseline` is the Linear SVM. Overall best results are in bold. IT = Italian; ES = Spanish; NL = Dutch; EN = English.



(a) Performance per dataset

(b) Performance per language

Figure 1: Average performances of the multilingual models for each class of the coarse-grain dataset.

5.1 Multilingual Models

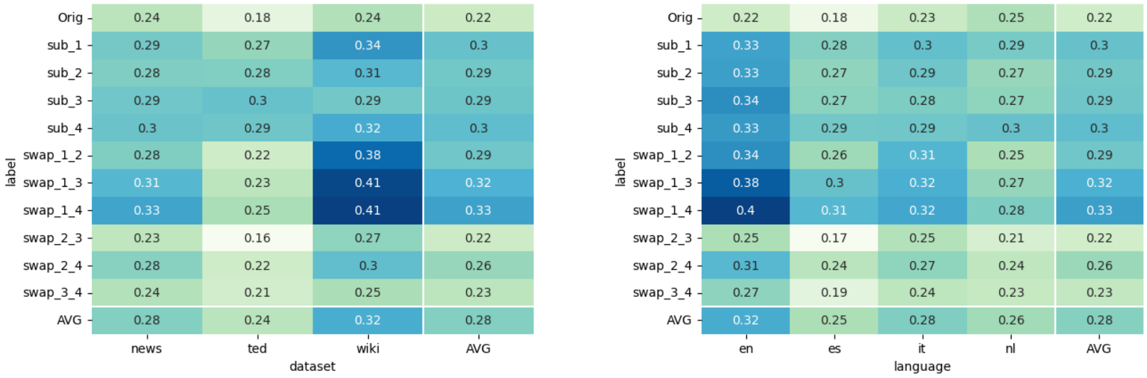
Coarse-grained Table 2, reports the models' accuracy in solving the three-labels classification task. All LMs perform significantly better than random prediction, while the baseline almost always aligns with it. With few exceptions, Wikipedia and News are easier to deal with than TED Talks. Unsurprisingly, the best results - across all genres and all languages - are obtained by `mDeBERTa-v3`. Figure 1 shows the LMs' performance by class. Figure 1a reports average performance across languages per dataset, while Figure 1b shows results by language across all datasets. The results clearly indicate that the easiest passages to identify are always the unperturbed ones (**Orig**), followed by those perturbed with **Swap**. On the contrary, passages perturbed with **Sub** are the most challenging to detect. This seems rather counter-intuitive as introducing sentences from different text passages is expected to have a major disruptive effect in breaking coherence than simply swapping the order of sentences belonging to the same local text passage. In terms of languages, there is no consistent behavior across language families. English is the easiest to handle, while Spanish proves to be the most difficult. Italian and Dutch are similarly challenging.

Fine-grained Table 3 reports the results on the 11-labesl classification. All LMs perform better than random, while the baseline fails in almost all cases. Yet, the fine-grained benchmark is highly challenging, with the highest Accuracy being 0.39 on the Italian Wikipedia, making 🥞**TEXT-CAKE** more difficult than previous benchmarks (Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Laban et al., 2021). In general, Wikipedia and News are easier to deal with than TED Talks. These results reveal an ideal coherence complexity continuum, where Wikpedia and TED Talks represent the two opposite extremes, suggesting a relationship between writing conventions (and style) and the expression of coherence. Again, the best results - across all genres and all languages - are obtained by `mDeBERTa-v3`. Figure 2 shows the LMs' average performance by class. Figure 2a reports average performance across languages per dataset, while Figure 2b shows results by language across all datasets. In contrast to the coarse-grained scenario, lack of coherence, especially in **Swap** perturbations, is easier to detect than its presence (**Orig**). The strongest coherence disruption occurs when the first sentence is swapped or substituted, breaking the logical order.

Overall, models with the same basic architec-

		Models			
Dataset	**Language**	`mBERT`	`XML-RoBERTa`	`mDeBERTa-v3`	`Baseline`
Wikipedia	IT	0.36	0.30	**0.39**	0.10
	ES	0.31	0.24	**0.32**	0.11
	NL	0.26	0.25	**0.37**	0.08
	EN	0.34	0.30	**0.38**	0.13
News	IT	0.24	0.21	**0.33**	0.11
	ES	0.12	0.26	**0.28**	0.08
	NL	0.22	0.25	**0.31**	0.08
	EN	0.32	**0.38**	**0.38**	0.08
TED Talks	IT	0.18	0.24	**0.26**	0.07
	ES	0.17	0.25	**0.28**	0.07
	NL	0.17	0.24	**0.25**	0.07
	EN	0.19	**0.29**	**0.29**	0.08

Table 3: Results of the multilingual LMs across datasets and languages. `Baseline` is the Linear SVM. Overall best results are in bold.



(a) Performance per dataset

(b) Performance per language

Figure 2: Average performances of the multilingual models for each class.

tures in terms of depth and width behave differently mostly due to the pre-training objective and the vocabulary size. Model's size (i.e., total number of parameters) seems to be less relevant than expected or assumed. No model has balanced pre-training data across all languages, yet the differences do not seem to be due to varying sizes of the pre-trainig materials. Interestingly, even with unbalanced pre-training data (e.g., Spanish having more data in Wiki-100 and CC-100 datasets than Italian and Dutch but performing worse), models perform similarly on specific perturbed classes. This suggests that inherent language-specific writing styles influence model performance.

5.2 Monolingual Models

The monolingual experiments allow us to explore in more detail the connection between model's size, depth, and width. Working only with English, we also use the fanfictions dataset to investigate the expression and detection of coherence in narrative texts. We report the results on two sets of experiments: in-domain and out-of-domain. This will allows us to investigate generalization capabilities of models (and their dependence from the model's properties) as well to gain better insights on the relationship between coherence and text genres. This section reports a detailed set of experiments on the fine-grained dataset. For completeness, the performance of monolingual models on the coarse-grained dataset are reported in Appendix A.

In-domain Results An overview of the results is presented in Table 4. While the Wikipedia and News subsets remain the easiest to process, the Fanfiction subset proves less challenging than TED Talks. As expected, all models outperform the baseline, and a clear pattern emerges linking model size and depth (i.e., number of parameters and layers) to performance. The `Large` model consistently outperforms all others by at least 10 points. We would generally expect a linear rela-

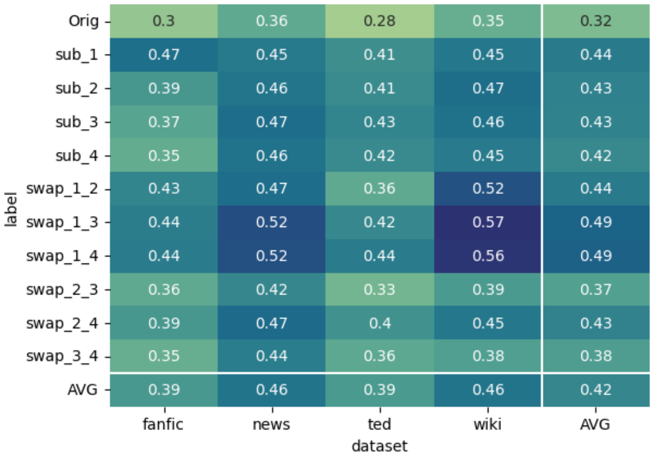


Figure 3: Overview of the results of the monolingual models on each class per dataset (in-domain). For each dataset the scores correspond to the average of all models.

	Models				
Dataset	XSmall	Small	Base	Large	BL
Wikipedia	0.41	0.37	0.46	**0.58**	0.13
News	0.41	0.39	0.46	**0.57**	0.08
TED Talks	0.35	0.31	0.39	**0.50**	0.08
Fanfiction	0.37	0.32	0.38	**0.49**	0.10
AVG	0.38	0.34	0.42	**0.53**	0.10

Table 4: In-domain results of the monolingual DeBERTa-v3 LM family on each dataset. BL corresponds to the Linear SVM. Best results are in bold.

tionship between model size, depth, and performance. However, this trend is only partially observed. While the Base model (the second largest) achieves the second-best performance, the XSmall model ranks third, outperforming the Small model, which shows the worst performance, consistent with findings from He et al. (2023). Scaling up models does improve performance, extending coherence detection to a broader range of downstream tasks impacted by larger models (Kaplan et al., 2020; Rae et al., 2021). The XSmall model outperforms the Small model by an average of 4 points, suggesting that model depth has a greater impact on coherence detection than model width.

Figure 3 shows the average Accuracy of the LMs for each dataset across all classes. Monolingual models outperform multilingual ones (Figure 2a), showing consistent trends in coherence modeling. For the Wikipedia, News, and TED Talks subsets, the behavior and patterns are the same as observed in the multilingual experiments, with Accuracy being higher. In the case of Fanfiction, the trend for **Swap** perturbations is similar to the other data portions, highlighting the importance of the first sentence in establishing coherence. For **Sub** perturbations, models perform worse when substituting all but the first sentence. Although the larger number of passages starting with the first sentence of a paragraph may have had an impact (see Table 1), this suggests that, in narratives, coherence is maintained throughout the text, unlike news articles where the ending sections are less coherent with the opening. As a result, the Accuracy for *Sub_4* and *Sub_3* is higher than for *Sub_1*.

Comparing TED Talks and Fanfiction highlights the impact of pre-training data. While these fanfictions were not used to train the DeBERTa models, similar text types (e.g., CC-STORIES and PG-19) were. This sensitivity to pre-training material, observed in multilingual experiments, emphasizes the need for the NLP community to document the data used to create LMs, thus facilitating the selection of promising models and reducing unnecessary experiments and resource consumption.

Out-of-domain Results To investigate generalization in out-of-domain settings, we finetuned all DeBERTa models using three data subsets for training and one left-out subset for testing. Table 5 shows the average scores for three finetuned models per dataset. The overall performance is similar to in-domain experiments, but the XSmall model, though narrow but with the same depth as the Base model, suffers the most (average Δ 0.115 against in-domain). On the contrary, the Base model (wider and nearly four times bigger), minimizes losses across all subsets.

	Models			
Dataset	XSmall	Small	Base	Large
Wikipedia	0.24 (Δ 0.17)	0.32 (Δ 0.05)	0.43 (Δ 0.03)	**0.53** (Δ 0.05)
News	0.33 (Δ 0.08)	0.29 (Δ 0.10)	0.43 (Δ 0.03)	**0.51** (Δ 0.06)
TED Talks	0.26 (Δ 0.09)	0.23 (Δ 0.08)	0.37 (Δ 0.02)	**0.45** (Δ 0.05)
Fanfiction	0.25 (Δ 0.12)	0.22 (Δ 0.10)	0.31 (Δ 0.07)	**0.38** (Δ 0.11)

Table 5: Out-of-domain results of the monolingual DeBERTa-v3 LMs' family on each dataset. Best results are in bold. In the parentheses, we report the deltas with respect to the in-domain results.

Genres have a less prominent impact on the performance drop. On average, the loss across genres ranges between 0.06 (News and TED Talks) and 0.1 (Fanfiction). Losses cannot be directly connected to the presence of comparable text genres in the pre-training. As a matter of fact, TED Talks, not present in the pre-trainng data of all LMs, has a lower drop that Fanfiction.

Generalization abilities are mostly affected by model's size. For instance, XSmall and Base have the same depth, but XSmall is a quarter the size of Base, leading to an average loss of 0.115 versus 0.03 for Base in out-of-domain experiments. Width may also contribute, as XSmall has half the attention heads of Base.

Classification Errors We further investigated the errors of the finetuned models, focusing on in-domain experiments. Detailed confusion matrices are in Appendix B.

Identifying the original text passage is the most challenging task, with frequent misclassifications as **Sub** perturbations, especially *Sub_1* and *Sub_4*. Errors rarely occur with **Swap** perturbations (all less than 1%). Models generally make errors across perturbed classes, particularly confusing adjacent sentences within **Sub** perturbations, such as *Sub_2* with *Sub_1*. **Swap** perturbations are often confused with **Sub** classes, like *Swap_1_2* with *Sub_1* or *Sub_2*. Distinct error patterns are not evident across models, suggesting minimal impact from size and parameter variations. All models struggle with *Swap_2_3* and *Swap_3_4* perturbed classes. Specific errors include DeBERTa-Small and DeBERTa-Base with *Swap_1_2*, and DeBERTa-XSmall with *Sub_4*. Error patterns are consistent across datasets, with no specific problematic cases per text genre.

6 Conclusions & Future Work

Coherence is a key textual property and essential for a variety of applications. We introduce 🥞**TEXT-CAKE**, a new benchmark and method to comprehensively investigate the sensitivity of LMs to local text coherence across genres and languages. Our method, potentially replicable for any language with diverse genre resources, reveals that this task poses a significant challenge for all examined LMs, underlining the importance of rigorously assessing LMs in discourse processing, especially given the advancements of generative models.

In our evaluation setting, we aimed to understand the connection between model architectures and their effectiveness when finetuned on this task. Not surprisingly, the granularity of the task has an impact on the models' performance. 🥞**TEXT-CAKE**, however, clearly indicates that local text coherence is more challenging to detect than at document level (Laban et al., 2021).

The multilingual experiments highlight the importance of pre-training objective and vocabulary size over the overall size of LMs. Additionally, the composition of pre-trained data significantly impacts performance, with text genre variation being more influential than language representations. Results from monolingual LMs, with fixed pre-training objectives and vocabulary size, highlight the impact of model depth and width on task-solving capabilities. Scaling-up proves effective in enhancing task performance, as larger models show better generalization also in a out-domain setting. When comparing models with varying depths, deeper models resulting more efficient for this specific task. This aligns with findings by Petty et al. (2023), who showed that while depth enhances generalization, its positive effects diminish if LM size and width remain fixed.

Future research should explore directly training language models for tasks related to local coherence detection to potentially achieve comparable or even better performance while bypassing the computational costs associated with fine-tuning.

Limitations

In this study, we focused exclusively on encoder-based models to investigate the effect of different parameters on coherence detection. This decision was primarily motivated by the need for consistency in the model architecture across different dimensions. We chose the `DeBERTa-v3` models because it maintains a consistent structure and pretraining objectives while being available in different shapes and sizes, allowing us to study the effect of model width and depth in isolation. In contrast, the available decoder-only or encoder-decoder models vary both dimensions simultaneously, which only allows us to study the effect of the number of parameters rather than the shape of the model. However, this exclusion is a limitation of our work, and future research could benefit from incorporating these models to provide a more comprehensive understanding of coherence detection across different model architectures.

The methodology used to create our dataset is language and genre independent but it requires sources with adequate size to ensure meaningful analysis. This requirement may bias the dataset towards certain types of sources that are more readily available or easily accessible in larger volumes. As a result, the findings of this study might not fully generalize to sources with limited availability or those not well-represented in the dataset. Addressing this limitation in future work could involve developing techniques to effectively utilize smaller or less conventional sources, thereby broadening the applicability of the research outcomes.

Acknowledgments

This work has been supported by:

- FAIR - Future AI Research (PE00000013) projects under the NRRP MUR program funded by the NextGenerationEU.;
- XAI-CARE - PNRRMAD-2022-12376692 projects under the NRRP MUR program funded by the NextGenerationEU;
- The project “Human in Neural Language Models” (IsC93_HiNLM), funded by CINECA3 under the ISCRA initiative;
- Language Of Dreams: the relationship between sleep mentation, neurophysiology, and neurological disorders - PRIN 2022 2022BNE97C_SH4_PRIN2022.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Dominique Brunato, Felice Dell’Orletta, Irene Dini, and Andrea Amelio Ravelli. 2023. Coherent or not? stressing a neural language model for discourse coherence in multiple languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10690–10700, Toronto, Canada. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China. Association for Computational Linguistics.

Robert-Alain De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*, volume 1. Longman London.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA. Association for Computational Linguistics.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Youmna Farag, Josef Valvoda, Helen Yannakoudakis, and Ted Briscoe. 2020. Analyzing neural discourse coherence models. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 102–112, Online. Association for Computational Linguistics.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece. Association for Computational Linguistics.

Mary Godunova and Ekaterina Voloshina. 2024. Probing of pretrained multilingual models on the knowledge of discourse. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 78–90, St. Julians, Malta. Association for Computational Linguistics.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? Event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman London.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. *CoRR*, abs/2104.05882.

Murathan Kurfalı and Robert Östling. 2021. Probing multilingual language models for discourse. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 8–19, Online. Association for Computational Linguistics.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020. Invisible to people but not to machines: Evaluation of style-aware headlinegeneration in absence of reliable human judgment. In *International Conference on Language Resources and Evaluation*.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4328–4339.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Panitan Muangkammuen, Sheng Xu, Fumiyo Fukumoto, Kanda Runapongsa Saikaew, and Jiyi Li. 2020. A neural local coherence analysis model for clarity text scoring. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2138–2143, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Conference on Computational Natural Language Learning*.

Alberto Parola, Jessica Mary Lin, Arndis Simonsen, Vibeke Bliksted, Yuan Zhou, Huiling Wang, Lana Inoue, Katja Koelkebeck, and Riccardo Fusaroli. 2023. Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of nlp automated measures of coherence. *Schizophrenia Research*, 259:59–70. Language and Speech Analysis in Schizophrenia and Related Psychoses.

Jackson Petty, Sjoerd van Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2023. The impact of depth and width on transformer language model generalization. *arXiv e-prints*, pages arXiv–2310.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González. 2022. DACSA: A large-scale dataset for automatic summarization of Catalan and Spanish newspaper articles. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931–5943, Seattle, United States. Association for Computational Linguistics.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640.

Sanna-Kaisa Tanskanen. 2006. *Collaborating towards coherence: Lexical cohesion in English discourse*. John Benjamins, Amsterdam/Philadelphia.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330, Vancouver, Canada. Association for Computational Linguistics.

Yuan Wang and Minghe Guo. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2):460.

Chia-Lun Yeh, Babak Loni, Mariëlle Hendriks, Henrike Reinhardt, and Anne Schuth. 2019. Dpgmedia2019: A dutch news dataset for partisanship detection.

A Monolingual experiments on coarse-grained dataset

Table A presents the performance of English monolingual models on the coarse-grained dataset. All models significantly outperform the SVM baseline. As expected, the `Large` model achieves the best results, followed by the `Base` model. Notably, on this relatively easier dataset, the performance difference between the `XSmall` and `Small` models is almost negligible. This suggests that in this case, the model's depth is less critical for solving the task, with the `Small` model nearly closing the performance gap due to its higher number of parameters.

	Models				
Dataset	`XSmall`	`Small`	`Base`	`Large`	`BL`
Wikipedia	0.61	0.60	0.66	**0.72**	0.35
News	0.61	0.61	0.68	**0.73**	0.35
TED Talks	0.56	0.55	0.62	**0.69**	0.34
Fanfiction	0.62	0.62	0.66	**0.68**	0.34
AVG	0.60	0.60	0.66	**0.71**	0.34

Table A: In-domain results of the monolingual `DeBERTa-v3` LM family on each coarse-grained dataset. `BL` corresponds to the Linear SVM. Best results are in bold.

B Results on classes

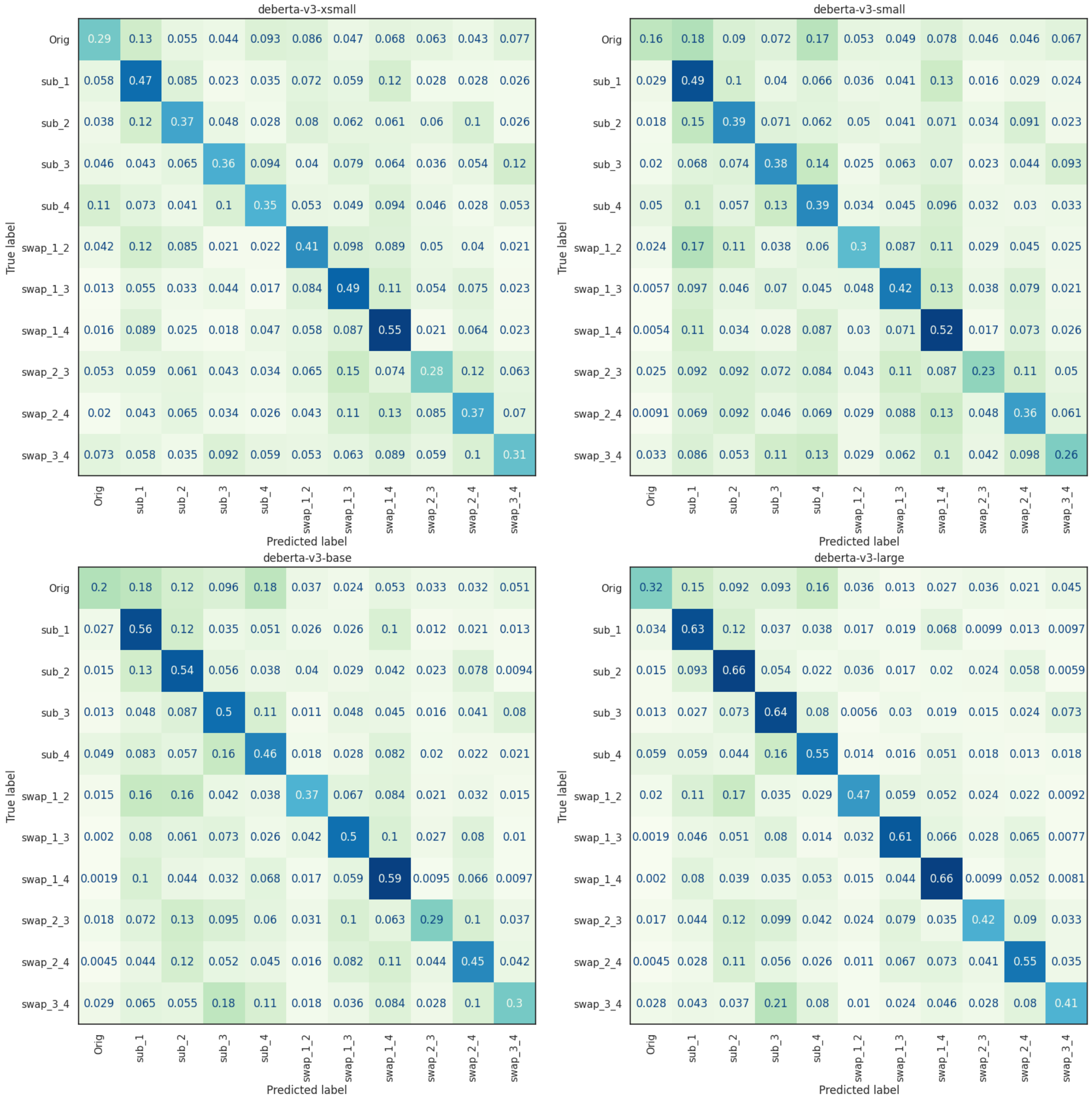
Figure A rdisplays the confusion matrices for the classification on the fine-grained dataset, organized by each monolingual model. Each confusion matrix represents the average values computed across all datasets. In contrast, Figure B presents the results for each dataset, averaged across all monolingual models.

C Model's parameters and sizes

Table B lists the number of parameters for all the models used in our experiments, as specified in the official documentation.

D Finetuning Hyperparamters

Table C presents the hyperparameters and GPUs utilized to fine-tune all the models during our experiments.



deberta-v3-xsmall

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.29	0.13	0.055	0.044	0.093	0.086	0.047	0.068	0.063	0.043	0.077
sub_1	0.058	0.47	0.085	0.023	0.035	0.072	0.059	0.12	0.028	0.028	0.026
sub_2	0.038	0.12	0.37	0.048	0.028	0.08	0.062	0.061	0.06	0.1	0.026
sub_3	0.046	0.043	0.065	0.36	0.094	0.04	0.079	0.064	0.036	0.054	0.12
sub_4	0.11	0.073	0.041	0.1	0.35	0.053	0.049	0.094	0.046	0.028	0.053
swap_1_2	0.042	0.12	0.085	0.021	0.022	0.41	0.098	0.089	0.05	0.04	0.021
swap_1_3	0.013	0.055	0.033	0.044	0.017	0.084	0.49	0.11	0.054	0.075	0.023
swap_1_4	0.016	0.089	0.025	0.018	0.047	0.058	0.087	0.55	0.021	0.064	0.023
swap_2_3	0.053	0.059	0.061	0.043	0.034	0.065	0.15	0.074	0.28	0.12	0.063
swap_2_4	0.02	0.043	0.065	0.034	0.026	0.043	0.11	0.13	0.085	0.37	0.07
swap_3_4	0.073	0.058	0.035	0.092	0.059	0.053	0.063	0.089	0.059	0.1	0.31

deberta-v3-small

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.16	0.18	0.09	0.072	0.17	0.053	0.049	0.078	0.046	0.046	0.067
sub_1	0.029	0.49	0.1	0.04	0.066	0.036	0.041	0.13	0.016	0.029	0.024
sub_2	0.018	0.15	0.39	0.071	0.062	0.05	0.041	0.071	0.034	0.091	0.023
sub_3	0.02	0.068	0.074	0.38	0.14	0.025	0.063	0.07	0.023	0.044	0.093
sub_4	0.05	0.1	0.057	0.13	0.39	0.034	0.045	0.096	0.032	0.03	0.033
swap_1_2	0.024	0.17	0.11	0.038	0.06	0.3	0.087	0.11	0.029	0.045	0.025
swap_1_3	0.0057	0.097	0.046	0.07	0.045	0.048	0.42	0.13	0.038	0.079	0.021
swap_1_4	0.0054	0.11	0.034	0.028	0.087	0.03	0.071	0.52	0.017	0.073	0.026
swap_2_3	0.025	0.092	0.092	0.072	0.084	0.043	0.11	0.087	0.23	0.11	0.05
swap_2_4	0.0091	0.069	0.092	0.046	0.069	0.029	0.088	0.13	0.048	0.36	0.061
swap_3_4	0.033	0.086	0.053	0.11	0.13	0.029	0.062	0.1	0.042	0.098	0.26

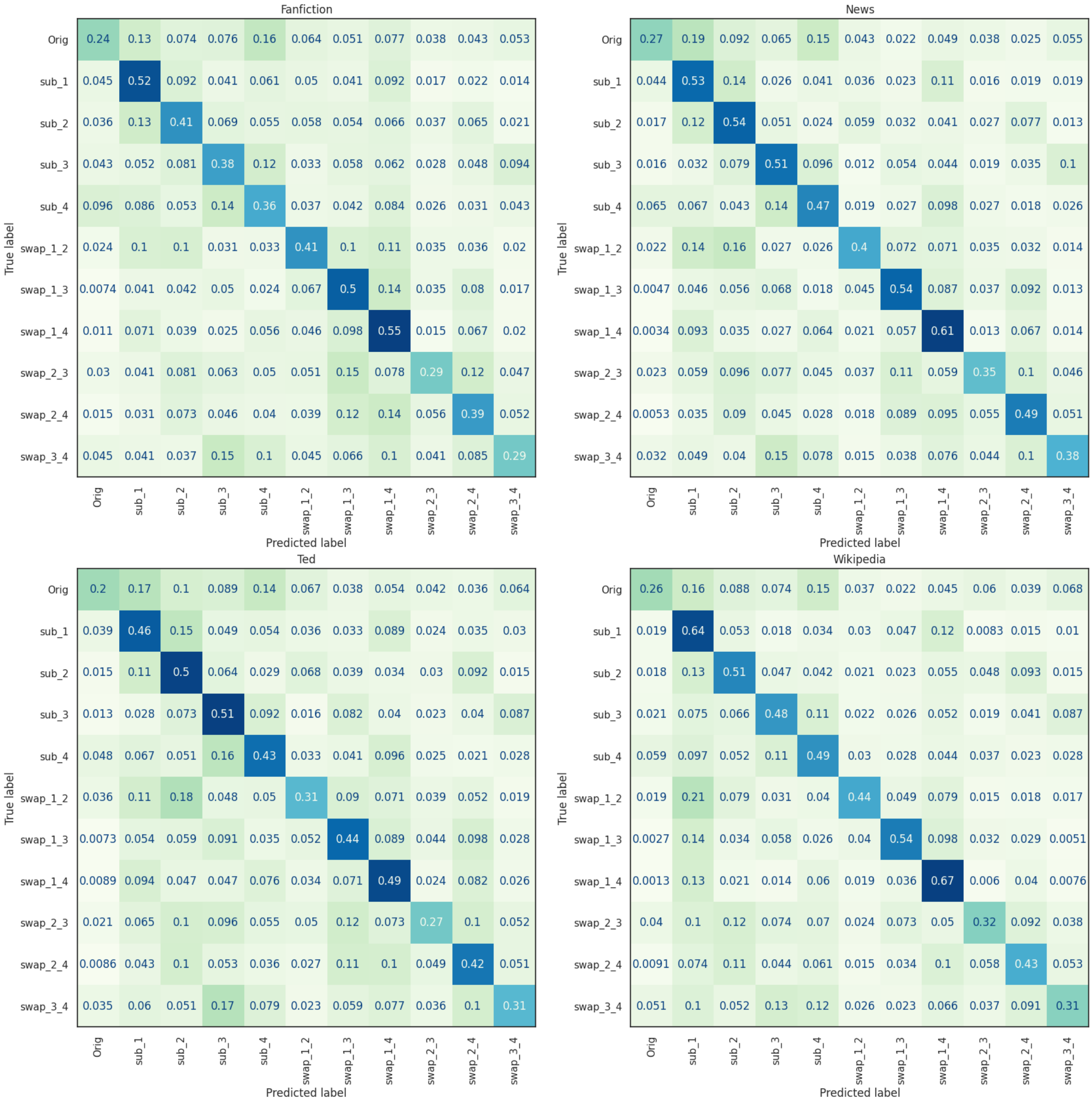
deberta-v3-base

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.2	0.18	0.12	0.096	0.18	0.037	0.024	0.053	0.033	0.032	0.051
sub_1	0.027	0.56	0.12	0.035	0.051	0.026	0.026	0.1	0.012	0.021	0.013
sub_2	0.015	0.13	0.54	0.056	0.038	0.04	0.029	0.042	0.023	0.078	0.0094
sub_3	0.013	0.048	0.087	0.5	0.11	0.011	0.048	0.045	0.016	0.041	0.08
sub_4	0.049	0.083	0.057	0.16	0.46	0.018	0.028	0.082	0.02	0.022	0.021
swap_1_2	0.015	0.16	0.16	0.042	0.038	0.37	0.067	0.084	0.021	0.032	0.015
swap_1_3	0.002	0.08	0.061	0.073	0.026	0.042	0.5	0.1	0.027	0.08	0.01
swap_1_4	0.0019	0.1	0.044	0.032	0.068	0.017	0.059	0.59	0.0095	0.066	0.0097
swap_2_3	0.018	0.072	0.13	0.095	0.06	0.031	0.1	0.063	0.29	0.1	0.037
swap_2_4	0.0045	0.044	0.12	0.052	0.045	0.016	0.082	0.11	0.044	0.45	0.042
swap_3_4	0.029	0.065	0.055	0.18	0.11	0.018	0.036	0.084	0.028	0.1	0.3

deberta-v3-large

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.32	0.15	0.092	0.093	0.16	0.036	0.013	0.027	0.036	0.021	0.045
sub_1	0.034	0.63	0.12	0.037	0.038	0.017	0.019	0.068	0.0099	0.013	0.0097
sub_2	0.015	0.093	0.66	0.054	0.022	0.036	0.017	0.02	0.024	0.058	0.0059
sub_3	0.013	0.027	0.073	0.64	0.08	0.0056	0.03	0.019	0.015	0.024	0.073
sub_4	0.059	0.059	0.044	0.16	0.55	0.014	0.016	0.051	0.018	0.013	0.018
swap_1_2	0.02	0.11	0.17	0.035	0.029	0.47	0.059	0.052	0.024	0.022	0.0092
swap_1_3	0.0019	0.046	0.051	0.08	0.014	0.032	0.61	0.066	0.028	0.065	0.0077
swap_1_4	0.002	0.08	0.039	0.035	0.053	0.015	0.044	0.66	0.0099	0.052	0.0081
swap_2_3	0.017	0.044	0.12	0.099	0.042	0.024	0.079	0.035	0.42	0.09	0.033
swap_2_4	0.0045	0.028	0.11	0.056	0.026	0.011	0.067	0.073	0.041	0.55	0.035
swap_3_4	0.028	0.043	0.037	0.21	0.08	0.01	0.024	0.046	0.028	0.08	0.41

Figure A: Confusion matrix for each model.



Fanfiction

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.24	0.13	0.074	0.076	0.16	0.064	0.051	0.077	0.038	0.043	0.053
sub_1	0.045	0.52	0.092	0.041	0.061	0.05	0.041	0.092	0.017	0.022	0.014
sub_2	0.036	0.13	0.41	0.069	0.055	0.058	0.054	0.066	0.037	0.065	0.021
sub_3	0.043	0.052	0.081	0.38	0.12	0.033	0.058	0.062	0.028	0.048	0.094
sub_4	0.096	0.086	0.053	0.14	0.36	0.037	0.042	0.084	0.026	0.031	0.043
swap_1_2	0.024	0.1	0.1	0.031	0.033	0.41	0.1	0.11	0.035	0.036	0.02
swap_1_3	0.0074	0.041	0.042	0.05	0.024	0.067	0.5	0.14	0.035	0.08	0.017
swap_1_4	0.011	0.071	0.039	0.025	0.056	0.046	0.098	0.55	0.015	0.067	0.02
swap_2_3	0.03	0.041	0.081	0.063	0.05	0.051	0.15	0.078	0.29	0.12	0.047
swap_2_4	0.015	0.031	0.073	0.046	0.04	0.039	0.12	0.14	0.056	0.39	0.052
swap_3_4	0.045	0.041	0.037	0.15	0.1	0.045	0.066	0.1	0.041	0.085	0.29

News

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.27	0.19	0.092	0.065	0.15	0.043	0.022	0.049	0.038	0.025	0.055
sub_1	0.044	0.53	0.14	0.026	0.041	0.036	0.023	0.11	0.016	0.019	0.019
sub_2	0.017	0.12	0.54	0.051	0.024	0.059	0.032	0.041	0.027	0.077	0.013
sub_3	0.016	0.032	0.079	0.51	0.096	0.012	0.054	0.044	0.019	0.035	0.1
sub_4	0.065	0.067	0.043	0.14	0.47	0.019	0.027	0.098	0.027	0.018	0.026
swap_1_2	0.022	0.14	0.16	0.027	0.026	0.4	0.072	0.071	0.035	0.032	0.014
swap_1_3	0.0047	0.046	0.056	0.068	0.018	0.045	0.54	0.087	0.037	0.092	0.013
swap_1_4	0.0034	0.093	0.035	0.027	0.064	0.021	0.057	0.61	0.013	0.067	0.014
swap_2_3	0.023	0.059	0.096	0.077	0.045	0.037	0.11	0.059	0.35	0.1	0.046
swap_2_4	0.0053	0.035	0.09	0.045	0.028	0.018	0.089	0.095	0.055	0.49	0.051
swap_3_4	0.032	0.049	0.04	0.15	0.078	0.015	0.038	0.076	0.044	0.1	0.38

Ted

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.2	0.17	0.1	0.089	0.14	0.067	0.038	0.054	0.042	0.036	0.064
sub_1	0.039	0.46	0.15	0.049	0.054	0.036	0.033	0.089	0.024	0.035	0.03
sub_2	0.015	0.11	0.5	0.064	0.029	0.068	0.039	0.034	0.03	0.092	0.015
sub_3	0.013	0.028	0.073	0.51	0.092	0.016	0.082	0.04	0.023	0.04	0.087
sub_4	0.048	0.067	0.051	0.16	0.43	0.033	0.041	0.096	0.025	0.021	0.028
swap_1_2	0.036	0.11	0.18	0.048	0.05	0.31	0.09	0.071	0.039	0.052	0.019
swap_1_3	0.0073	0.054	0.059	0.091	0.035	0.052	0.44	0.089	0.044	0.098	0.028
swap_1_4	0.0089	0.094	0.047	0.047	0.076	0.034	0.071	0.49	0.024	0.082	0.026
swap_2_3	0.021	0.065	0.1	0.096	0.055	0.05	0.12	0.073	0.27	0.1	0.052
swap_2_4	0.0086	0.043	0.1	0.053	0.036	0.027	0.11	0.1	0.049	0.42	0.051
swap_3_4	0.035	0.06	0.051	0.17	0.079	0.023	0.059	0.077	0.036	0.1	0.31

Wikipedia

	Orig	sub_1	sub_2	sub_3	sub_4	swap_1_2	swap_1_3	swap_1_4	swap_2_3	swap_2_4	swap_3_4
Orig	0.26	0.16	0.088	0.074	0.15	0.037	0.022	0.045	0.06	0.039	0.068
sub_1	0.019	0.64	0.053	0.018	0.034	0.03	0.047	0.12	0.0083	0.015	0.01
sub_2	0.018	0.13	0.51	0.047	0.042	0.021	0.023	0.055	0.048	0.093	0.015
sub_3	0.021	0.075	0.066	0.48	0.11	0.022	0.026	0.052	0.019	0.041	0.087
sub_4	0.059	0.097	0.052	0.11	0.49	0.03	0.028	0.044	0.037	0.023	0.028
swap_1_2	0.019	0.21	0.079	0.031	0.04	0.44	0.049	0.079	0.015	0.018	0.017
swap_1_3	0.0027	0.14	0.034	0.058	0.026	0.04	0.54	0.098	0.032	0.029	0.0051
swap_1_4	0.0013	0.13	0.021	0.014	0.06	0.019	0.036	0.67	0.006	0.04	0.0076
swap_2_3	0.04	0.1	0.12	0.074	0.07	0.024	0.073	0.05	0.32	0.092	0.038
swap_2_4	0.0091	0.074	0.11	0.044	0.061	0.015	0.034	0.1	0.058	0.43	0.053
swap_3_4	0.051	0.1	0.052	0.13	0.12	0.026	0.023	0.066	0.037	0.091	0.31

Figure B: Confusion matrix for each dataset.

	Models						
	Multilingual			Monolingual			
Parameter	mBERT	XML-RoBERTa	mDeBERTa-v3	XSmall	Small	Base	Large
Number of Layers	12	12	12	12	6	12	24
Hidden size	768	768	768	384	768	768	1,024
FNN inner hidden size	3,072	3,072	3,072	1,536	3,072	3,072	4,096
Attention Heads	12	12	12	6	12	12	12
Attention Head size	64	64	64	64	64	64	64
Vocabulary Size	30.5k	30.5K	250K	128K	128K	128K	128K
Total Parameters	110M	125M	86M	22M	44M	86M	304M

Table B: Summary overview of the hyperparamters of the models in analysis. All multilingual models refer to their base versions. The names of the monolingual models indicates different versions of DeBERTa-v3.

	Models						
	Multilingual			Monolingual			
Hyperparameter	mBERT	XML-RoBERTa	mDeBERTa-v3	XSmall	Small	Base	Large
learning rate	$2e-5$	$5e-6$	$2e-5$	$2e-5$	$2e-5$	$2e-5$	$8e-6$
Epochs	10	10	10	10	10	10	10
Batch size	64	64	64	64	64	64	64
Warm-up ratio	0.1	0.1	0.1	0.1	0.1	0.1	0.1
GPU							
1 NVIDIA GeForce RTX 2080 Ti (12GB)	✓	✓	✓	✓	✓	✓	–
2 NVIDIA GeForce RTX 4090 (24GB)	–	–	–	–	–	–	✓

Table C: Overview of the hyperparamters and GPUs used to finetuned all models. All multilingual models refer to their base versions. The names of the monolingual models indicates different versions of DeBERTa-v3.