An Information-Theoretic Approach to Analyze NLP Classification Tasks

Luran Wang, Mark Gales, Vatsal Raina
ALTA Institute, University of Cambridge
{lw703,mjfg,vr311}@cam.ac.uk

Abstract

Understanding the contribution of the inputs on the output is useful across many tasks. This work provides an information-theoretic framework to analyse the influence of inputs for text classification tasks. Natural language processing (NLP) tasks take either a single or multiple text elements to predict an output variable. Each text element has two components: the semantic meaning and a linguistic realization. Multiple-choice reading comprehension (MCRC) and sentiment classification (SC) are selected to showcase the framework. For MCRC, it is found that the relative context influence on the output reduces on more challenging datasets. In particular, more challenging contexts allows greater variation in the question complexity. Hence, test creators need to carefully consider the choice of the context when designing multiple-choice questions for assessment. For SC, it is found the semantic meaning of the input dominates compared to its linguistic realization when determining the sentiment. The framework is made available at: https://github.com/WangLuran/nlp-element-influence.

1 Introduction

Natural Language Processing (NLP) requires machines to understand language to perform a specific task (Chowdhary and Chowdhary, 2020). NLP tasks take a single (e.g. summarization (Widyassari et al., 2022), sentiment classification (Wankhade et al., 2022), machine translation (Stahlberg, 2020)) or multiple (e.g. reading comprehension (Baradaran et al., 2022), question generation (Kurdi et al., 2020)) text elements at the input and return a specific output. Each input text element can further be partitioned into its semantic content and the linguistic realization. Semantic refers to the inherent meaning while the linguistic realization is the specific wording to present the meaning. There are several possible linguistic realizations for any semantic content. Therefore, for all NLP tasks, the output variable has contributions from at least two components: the semantic meaning of the element and the specific linguistic realization. Here, *element* refers to a specific input that is formed of exactly two *component*s.

We analyze the relative sensitivity of the output variable to each of the input elements as well as in terms of the breakdown between the elemental semantic content and its corresponding linguistic realization. A theoretical information-theoretic approach is applied to find the shared information content between each input component and the output variable. Here, the information-theoretic approach is framed for NLP classification tasks where the set of input components influence the output probability distribution over a discrete set of classes. We select multiple-choice reading comprehension (MCRC) and sentiment classification (SC) as case studies for the analysis.

MCRC requires the correct answer option to be selected based on several input elements: the context paragraph, the question and the set of answer options. Multiple-choice (MC) assessments are a widely employed method for evaluating the competencies of candidates across diverse settings and tasks on a global scale (Lai et al., 2017a; Richardson et al., 2013a; Sun et al., 2019; Levesque et al., 2012). Given their consequential impact on real-world decisions, the selection of appropriate MC questions tailored to specific scenarios is important for content creators. Consequently, there is a need to comprehend the underlying factors that contribute to the complexity of these assessments.

Complexity of an MC question is best modelled by the distribution over the answer options by human test takers. Therefore, by understanding the influence of each input element on the output distribution, content creators can be better informed to what extent the complexity of an MC question can be controlled from changing each of the input ele-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 530–551

ments. Moreover, analyzing the contribution of the semantic content vs the linguistic realization on the output human distribution informs the impact of the specific word choice in the element on the question complexity. However, it is not scalable to measure the variation in the output human distribution with variation in each of the input elements. Liusie et al. (2023c) demonstrated that the output distribution of automated systems is aligned (with minimal re-shaping parameters) to the human distribution. Therefore, the information-theoretic framework is applied to the output probability distribution by an automated comprehension system.

SC is a common NLP classification task where the dominant sentiment class must be selected from a discrete set of sentiments based on a block of input text. This is an example of a single input text element NLP task. The information-theoretic approach is applied here to understand the role of the semantic content and the linguistic realization on the output distribution over the sentiment classes for popular datasets. It is interesting to analyze SC as ideally the sentiment of a text block should be based on only its semantic meaning. Here, we determine whether this ideal is held in practice.

2 Related Work

Features or variables are separate properties that are input to tabular machine learning models to predict a target variable (Hwang and Song, 2023). Feature importance is an active area of research (Huang et al., 2023) where the influence of each feature on the output is determined. The ability to determine which features are most important is useful across many verticals e.g. computer assisted medical diagnosis (Rudin, 2019), weather forecasting (Malinin et al., 2021), fraud detection (Xu et al., 2023) and customer churn prediction (Al-Shourbaji et al., 2023). Similarly, we explore the importance of different aspects (can be interpreted as features) at the input including individual elements and the semantic vs linguistic components for NLP text classification tasks. Typically, the structured nature of tabular data allows common feature selection algorithms to be applied including LASSO (Tibshirani, 1996), marginal screening (Fan and Lv, 2008), orthogonal matching pursuit (Pati et al., 1993) and decision tree based (Costa and Pedreira, 2023). Due to the relatively unstructured nature of text data (compared to tabular data), we propose an information-theoretic approach to identify the most influential inputs.

Sugawara et al. (2017) find a weak correlation between question difficulty and context readability for MCRC. Additionally, Sugawara et al. (2020) consider the impact on MCRC datasets when input elements are omitted. We propose instead an automated information-theoretic framework for this analysis. Finally, Sorensen et al. (2022) apply an information-theoretic approach for prompt engineering. Our approach can instead be generalized to any NLP classification task.

3 Theory

Here, we describe the generalized framework to analyze the influence of different elements in NLP text classification tasks: 1. the individual influence of each input element on the output class distribution; 2. the contribution of the semantic content vs its linguistic realization component for a given element. Let an NLP task consist of a set of elements, $\{x_1, \ldots, x_N\} = \mathbf{x}$ and the output, y, such that:

$$P(y) = \mathbb{E}_{P(\mathbf{x})} P\left(y|\mathbf{x}\right) \tag{1}$$

Let $\mathbf{X}$ denote the random variables of each the corresponding instances $\mathbf{x}$. Similarly, let Y be the random variable for an instance of the output, y.

To measure the influence of input $\mathbf{x}$ on output y, a good metric is the mutual information (Depeweg et al., 2018; Malinin and Gales, 2018) which measures how the output changes due to variation in the input. Thus we can define $\mathcal{I}(Y;\mathbf{X})$ a measure of the total input influence. Similarly we can define the influence from an individual element, X_j, $\mathcal{I}(Y;X_j)$ and it should obey:

$$\underbrace{\mathcal{I}(Y;X_j)}_{\text{element}} = \underbrace{\mathcal{I}\left(Y;\mathbf{X}\right)}_{\text{total}} - \underbrace{\mathcal{I}\left(Y;\mathbf{X}\backslash X_j|X_j\right)}_{\text{other}} \tag{2}$$

For each element X_j, its influence is always determined by two components: $X_j^{(s)}$, the semantic information and a relating linguistic realization method which turns an abstract meaning into natural language. Thus, we can calculate the semantic influence as $\mathcal{I}(Y;X_j^{(s)})$ and the linguistic influence implicitly $\mathcal{I}(Y;X_j|X_j^{(s)})$. They should satisfy:

$$\underbrace{\mathcal{I}(Y;X_j)}_{\text{element}} = \underbrace{\mathcal{I}\left(Y;X_j^{(s)}\right)}_{\text{semantic}} + \underbrace{\mathcal{I}\left(Y;X_j|X_j^{(s)}\right)}_{\text{linguistic}} \tag{3}$$

In practice for an element, x_j, its semantic content is too abstract to be available. Instead we get access

to one of its realization $\tilde{r}_j$ which is considered to be generated from its unobserved semantic content, $x_j^{(s)}$. A set of possible realizations of this semantic element, $\mathcal{R}^{(i)}$, are additionally where each member of this set is, $r_j^{(i)}$ drawn as

$$r_j^{(i)} \sim P_r(r|\tilde{r}_i) \approx P_r(r|x_i^{(s)}) \tag{4}$$

With these settings, the mutual information is calculated as follows. The total influence is:

$$\begin{aligned} &\mathcal{I}(Y;\mathbf{X}) \\ &= \mathcal{H}\left(\mathbb{E}_{P(\mathbf{x})}[P(y|\mathbf{x})]\right) - \mathbb{E}_{P(\mathbf{x})}[\mathcal{H}(P(y|\mathbf{x}))] \end{aligned} \tag{5}$$

We can also get the element influence as :

$$\begin{aligned} &\mathcal{I}(Y;X_j) \\ &= \mathcal{H}\left(\mathbb{E}_{P(\mathbf{x})}[P(y|\mathbf{x})]\right) - \mathbb{E}_{P(x_j)}\left[\mathcal{H}(P(y|x_j))\right] \end{aligned} \tag{6}$$

It can be decomposed as the semantic influence:

$$\begin{aligned} \mathcal{I}\left(Y;X_j^{(s)}\right) = \mathcal{H}\left(\mathbb{E}_{P(\mathbf{x})}[P(y|\mathbf{x})]\right) \\ -\mathbb{E}_{P\left(x_j^{(s)}\right)}\left[\mathcal{H}\left(P\left(y|x_j^{(s)}\right)\right)\right] \end{aligned} \tag{7}$$

and the linguistic influence:

$$\begin{aligned} \mathcal{I}\left(Y;X_j|X_j^{(s)}\right) = \mathbb{E}_{P\left(x_j^{(s)}\right)}\left[\mathcal{H}\left(P\left(y|x_j^{(s)}\right)\right)\right] \\ -\mathbb{E}_{P(x_j)}\left[\mathcal{H}\left(P(y|x_j)\right)\right] \end{aligned} \tag{8}$$

The relative contribution of an element to the total influence and of the semantic component for an element can respectively be expressed as:

$$\text{relative element influence} = \frac{\mathcal{I}(Y;X_j)}{\mathcal{I}(Y;\mathbf{X})} \tag{9}$$

$$\text{relative semantic influence} = \frac{\mathcal{I}\left(Y;X_j^{(s)}\right)}{\mathcal{I}(Y;X_j)} \tag{10}$$

3.1 Multiple-choice reading comprehension

In this task, candidates are provided with a context passage, c and a corresponding question, q. The objective is to determine the correct answer from a defined set of options, denoted as o. This process involves understanding the question and utilizing the context passage as a source of information to ascertain the most appropriate answer option. The output distribution can be categorised as:

$$P(y) = \mathbb{E}_{P(c,q,o)}P(y|c,q,o) \tag{11}$$

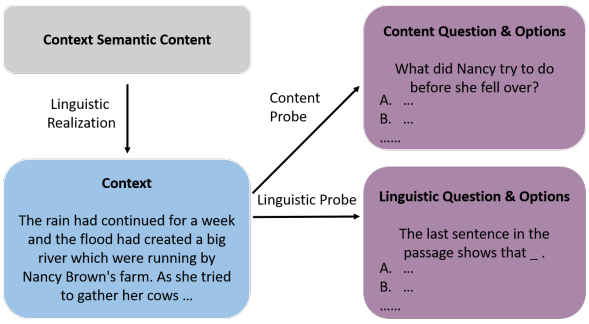


Figure 1: Data generation for the multiple-choice reading comprehension task.

3.1.1 Data generation

For a typical MCRC dataset, the data generation process is shown in Figure 1. A specific semantic content $c^{(s)}$ is chosen and a context c is generated when a certain linguistic realization is applied. Therefore, the influence of the context C can be divided into $\mathcal{I}(Y;C^{(s)}), \mathcal{I}(Y;C|C^{(s)})$ respectively. A similar procedure is applied on the questions and the options but usually (for the scope of this work) they are generated together as a question-option pair q: from a certain context, a content probe or linguistic probe is generated and then a question-option pair in natural language is a linguistic realization of this probe as described by: $P(q|c)$. Thus, the output distribution can be rewritten as:

$$P(y) = \mathbb{E}_{P(c^{(s)})}\mathbb{E}_{P(c|c^{(s)})}\mathbb{E}_{P(q|c)}P(y|q,c) \tag{12}$$

We consider only the questions generated from the semantic contents and ignore the questions constrained to a specific realization, by filtering out all questions generated from the specific linguistic realization of the context. This allows our investigation on the question influence to be agnostic of the original context realization.

3.1.2 Measure of component influence

The question-option pair in Equation 12 appear as $P(q|c)$, thus instead of $\mathcal{I}(Y;Q)$, we consider $\mathcal{I}(Y;Q|C)$ and get the decomposition:

$$\underbrace{\mathcal{I}(Y;C)}_{\text{context}} = \underbrace{\mathcal{I}(Y;C,Q)}_{\text{total}} - \underbrace{\mathcal{I}(Y;Q|C)}_{\text{question}} \tag{13}$$

The context influence can be further decomposed:

$$\underbrace{\mathcal{I}(Y;C)}_{\text{context}} = \underbrace{\mathcal{I}(Y;C^{(s)})}_{\text{semantic}} + \underbrace{\mathcal{I}(Y;C|C^{(s)})}_{\text{linguistic}} \tag{14}$$

Equations 13 and 14 are examples of Equations 2 and 3 respectively. Thus, similar to Section 3,

the influence terms can be calculated according to Equations 5 to 8. Besides the assumption made in Equation 4 which is general for all the tasks, a further assumption about the questions are made for the MCRC task: instead of sampling from the ideal question generation process, for the i^{th} context realization in the dataset, $\tilde{r}_i$, we only observe the question-option pairs generated by humans, $\tilde{\mathcal{Q}}^{(i)}$, where each member of this set is, $\tilde{q}_j^{(i)}$ drawn as:

$$\tilde{q}_j^{(i)} \sim P_{\text{man}}(q|\tilde{r}_i) \approx P_q\left(q|c_i^{(s)}\right) \tag{15}$$

If we generate several paraphrases conditional on the original context such that $r \sim P_{\text{gpt}}(r|c)$, we then make the following approximations:

$$\mathbb{E}_{P(c,q)}\left[P(y|c,q)\right] \approx \tag{16}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\frac{1}{|\mathcal{R}^{(i)}||\tilde{\mathcal{Q}}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)},\tilde{q}\in\tilde{\mathcal{Q}}^{(i)}}P(y|\tilde{q},r)$$

$$\mathbb{E}_{P(c,q)}\left[\mathcal{H}(P(y|c,q))\right] \approx \tag{17}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\frac{1}{|\mathcal{R}^{(i)}||\tilde{\mathcal{Q}}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)},\tilde{q}\in\tilde{\mathcal{Q}}^{(i)}}\mathcal{H}(P(y|\tilde{q},r))$$

$$\mathbb{E}_{P(c)}\left[\mathcal{H}(P(y|c))\right] \approx \tag{18}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\frac{1}{|\mathcal{R}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)}}\mathcal{H}(\frac{1}{|\tilde{\mathcal{Q}}^{(i)}|}\sum_{\tilde{q}\in\tilde{\mathcal{Q}}^{(i)}}P(y|\tilde{q},r))$$

$$\mathbb{E}_{P(c^{(s)})}\left[\mathcal{H}(P(y|c^{(s)}))\right] \approx \tag{19}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\mathcal{H}(\frac{1}{|\mathcal{R}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)}}\frac{1}{|\tilde{\mathcal{Q}}^{(i)}|}\sum_{\tilde{q}\in\tilde{\mathcal{Q}}^{(i)}}P(y|\tilde{q},r))$$

with n_s as the number of contexts in a dataset. The detailed derivation of Equations 16 to 19 are shown in Appendix A.

3.2 Sentiment classification

For the SC task, the candidate receives a sentence or a short paragraph x and then is requested to choose the sentiment class. Here we are only interested in the influence to the output y from semantic content $x^{(s)}$ and its linguistic realization method: $\mathcal{I}(Y;X^{(s)}), \mathcal{I}(Y;X|X^{(s)})$, as there is only one element at the input. Following Equation 3, the semantic and linguistic breakdown is expressed as:

$$\underbrace{\mathcal{I}(Y;X)}_{\text{text}} = \underbrace{\mathcal{I}(Y;X^{(s)})}_{\text{semantic}} + \underbrace{\mathcal{I}(Y;X|X^{(s)})}_{\text{linguistic}} \tag{20}$$

In practice, the following approximations are made:

$$\mathbb{E}_{P(x)}\left[P(y|x)\right] \approx \tag{21}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\frac{1}{|\mathcal{R}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)}}P(y|r)$$

$$\mathbb{E}_{P(x)}\left[\mathcal{H}(P(y|x))\right] \approx \tag{22}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\frac{1}{|\mathcal{R}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)}}\mathcal{H}(P(y|r))$$

$$\mathbb{E}_{P(x^{(s)})}\left[\mathcal{H}(P(y|x^{(s)}))\right] \approx \tag{23}$$
$$\frac{1}{n_{\mathbf{s}}}\sum_{i=1}^{n_{\mathbf{s}}}\mathcal{H}(\frac{1}{|\mathcal{R}^{(i)}|}\sum_{r\in\mathcal{R}^{(i)}}P(y|r))$$

4 Systems

4.1 Linguistic realization

To analyze the impact of the linguistic realization of a given text element, it is necessary to fix the semantic content of the element. In other work (such as Sugawara et al. (2022)), they attempt to evaluate the effect of linguistic content of the context for MCRC. However, they ignore the requirement to fix the semantic content. We employ a paraphrasing system to generate different linguistic realizations for the same semantic content of a text element. The paraphrasing approach is applied to the context in MCRC and to the input element in SC. To consider a broad range of linguistic realizations for a specific text's semantic content, we generate 8 paraphrases at different readability levels. Hence, we assume (this assumption is assessed in Appendix F) that the linguistic realizations at different readability levels maintain the same semantic content. To change the readability of the text element, we use the zero-shot method as in Farajidizaji et al. (2023) based on Equation 4 to generate the paraphrase with the j^{th} readability from the i^{th} context:

$$r_j^{(i)} \sim P_{\text{LLM}}(r|\tilde{r}_i) \tag{24}$$

In practice, the zero-shot large language model (LLM) (GPT-3.5-turbo) is fed with the original text along with an instruction to alter the language of the text to match the desired readability level. The model, not previously trained on this specific task, uses its pre-existing knowledge and understanding

of language structure to alter the readability whilst maintaining the same semantic meaning. The readability level is measured by Flesch reading-ease (Flesch, 1948) score (FRES). See the prompts in Appendix Table 7.

4.2 Reading comprehension

In this work, MCRC systems are required to return a probability distribution over the answer options. Two alternative architectures are considered for performing the reading comprehension task: encoder-only and decoder-only as shown in Figure 2.

Encoder-only is based on the works of Yu et al. (2020); Raina and Gales (2022a); Liusie et al. (2023b); Raina et al. (2023b). Within the family of encoder-only models, we consider a BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) based systems. Each option is individually encoded along with both the question and the context to produce a score. Softmax is then applied to the scores linked to each option, transforming them into a probability distribution. During inference, the anticipated answer is chosen as the option with the highest associated probability.

Inspired by Liusie et al. (2023a) and the recent success observed in finetuning large open-source instruction finetuned language models (Touvron et al., 2023a,b; Jiang et al., 2023, 2024; Tunstall et al., 2023) on various NLP tasks, this work additionally finetunes Llama-2 (Touvron et al., 2023b) as a decoder-only architecture. The context, question and answer options are concatenated into a single natural language prompt. As an autoregressive language model, Llama-2 is requested to effectively return a single token at the output, represented by a single logit distribution over the token vocabulary. The logits associated with the tokens *A,B,C,D* are respectively normalized using softmax to return the desired probability distribution over the answer options. As with the encoder-only architecture, the option with the highest probability is selected as the answer at inference time. All model outputs are calibrated post-hoc (see Appendix D.3).

4.3 Data complexity classification

standard	$c + q + o_A + o_B + o_C + o_D$
context	c
context-question	$c + q$

Table 1: Input formats for the complexity system.

Here, an automated complexity system takes all the components of an MC question and classifies it into one of the 3 classes: *easy*, *medium* or *hard*. The standard architecture is used for MC question complexity classification (Raina and Gales, 2022b; Benedetto, 2023) (see Appendix Figure 6). All elements of an MC question and concatenated together and fed into a transformer. The embedding of the prepended `[CLS]` token is taken as the sentence embedding, which is passed to a classification head, to return a distribution over the three complexity levels. To empirically investigate the relative importance of each element, various input formats are trialled. Table 1 presents different combinations of the context, question and answer options.

4.4 Sentiment classification

SC models take the input text and return a probability distribution over the set of sentiments. Here, the sentiments considered are {negative, positive }. We take the standard approach of taking a pretrained transformer encoder model (Vaswani et al., 2017) with a classification head at the output (Liusie et al., 2022). Similar to the encoder-only approach for MCRC and the data complexity classification system, the SC system only passes the hidden embedding of the `[CLS]` token to the classification head. Softmax normalizes the logits over the sentiments.

5 Experiments

5.1 Data

We use the RACE++ MCRC dataset (Lai et al., 2017b; Liang et al., 2019a) train split for training both the MC data complexity system and the MCRC model. RACE++ is the largest publicly available dataset from English exams in China partitioned into three difficulty levels: middle school, high school and college (see Appendix E.1 for details about the splits). Additionally, various MCRC datasets are considered as test sets for investigating the influence of each element including the test sets from RACE++, MCTest (Richardson et al., 2013b) and CMCQRD (Mullooly et al., 2023). MCTest requires machines to answer MCRC questions about fictional stories. CMCQRD is a small-scale MCRC dataset from the pre-testing stage partitioned into grade levels B1 to C2 on the Common European Framework of Reference for Languages scale.

For SC, we use IMDb (Maas et al., 2011), Yelp-polarity (Yelp) (Zhang et al., 2015) and Amazon-polarity (Amazon) (McAuley and Leskovec, 2013) datasets. IMDb has reviews from the Internet

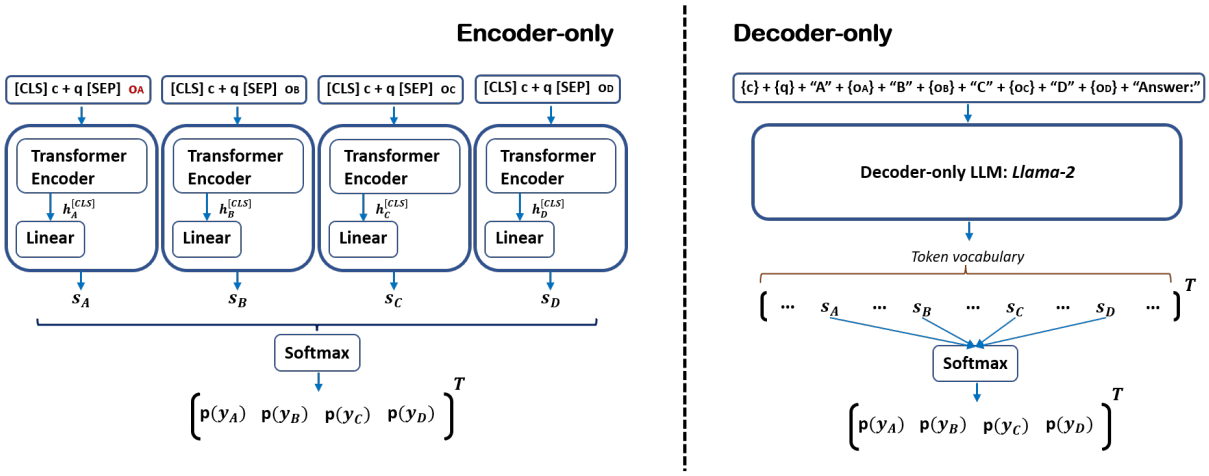


Figure 2: Architectures for multiple-choice reading comprehension with context, c, question, q and options, o.

		# examples	# options	# words	# questions	semantic diversity	linguistic diversity
MCRC	MCTest	142	4	209	4	$0.079_{\pm 0.015}$	$0.018_{\pm 0.006}$
	RACE++	1,007	4	278	3.7	$0.101_{\pm 0.015}$	$0.016_{\pm 0.007}$
	CMCQRD	150	4	683	5.5	$0.092_{\pm 0.010}$	$0.022_{\pm 0.011}$
SC	IMDB	500	2	226	-	$0.084_{\pm 0.019}$	$0.023_{\pm 0.011}$
	Yelp	500	2	133	-	$0.108_{\pm 0.037}$	$0.030_{\pm 0.024}$
	Amazon	500	2	74	-	$0.135_{\pm 0.024}$	$0.037_{\pm 0.022}$

Table 2: Statistics for multiple-choice reading comprehension (MCRC) and sentiment classification (SC) test datasets. See Appendix Table 5 for additional datasets.

Movie Database. Yelp consists of reviews where 1 or 2 stars is interpreted as negative while 4 or 5 stars is interpreted as positive. Amazon has reviews over a period of 13 years on various products. Hence, all 3 datasets are binary classification tasks.

Table 2 details the main statistics. All the MCRC test sets have 4 options while the selected SC test sets have 2 options: negative and positive. The number of words for MCRC is the lengths of the contexts. It is seen that the test sets have varying lengths from 200 to 700 words and 75 to 230 words for MCRC and SC tasks respectively. For the MCRC datasets, the total examples are reported after filtering out all linguistic probe questions (Appendix E.3 for the procedure). So the focus is only on the semantic probe questions as assumed in Section 3.1.2. For the SC test sets, a subset of 500 examples is selected for each dataset to remain within the financial budget for use of ChatGPT for the generation of different linguistic realizations.

The semantic diversity is also calculated for each dataset. This score is the mean cosine distance between each text embedding to the centroid of all embeddings (Raina et al., 2023a). Greater the score, greater the semantic variation in the set of texts being considered. The sentence embedder from Ni et al. (2022) is used to generate the embeddings[1]. The semantic diversity is calculated on the contexts for MCRC. Finally, the linguistic diversity calculates the mean variation in the embedding space for different linguistic realizations for each text.

5.2 Model details

The decoder-only implementation of the MCRC system is based upon the pretrained instruction finetuned Llama2-7B model[2]. The main paper for MCRC focuses only on the decoder-only implementation (see Appendix E.2 for the encoder-only implementations). ELECTRA-base (Clark et al., 2020) is selected for the data complexity evaluator models. The pretrained model is finetuned on the RACE++ train split with the complexity class (easy, medium or hard) as the label (hyperparameter tuning details in Appendix E.2). For SC, the main paper reports results based on a RoBERTa architecture. The selected model has been fine-

[1]Available at: https://huggingface.co/sentence-transformers/sentence-t5-base

[2]Available at https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

dataset	accuracy		influence				
	original	para	total	question	context	context-semantic	context-linguistic
MCTest	92.5	85.8	0.212	0.116 (54.7%)	0.096 (45.3%)	0.068 (70.6%)	0.028 (29.4%)
RACE++	86.0	82.9	0.298	0.161 (56.1%)	0.131 (43.9%)	0.108 (82.5%)	0.023 (17.5%)
CMCQRD	79.9	69.4	0.290	0.211 (72.7%)	0.079 (27.3%)	0.067 (83.8%)	0.012 (16.2%)

Table 3: Decomposition of total input influence for Llama-2 on MCRC datasets.

tuned on IMDb training data[3]. Due to the similarity in content between Yelp, Amazon and IMDb, the RoBERTa model finetuned on IMDb is also applied on all SC test sets. For reproducibility, see BERT (Devlin et al., 2018) system in Appendix E.2.

6 Results

6.1 Reading Comprehension

Table 3 presents the performance of Llama-2 on the various MCRC datasets. The highest accuracy is observed on MCTest with 92.5% and the lowest on CMCQRD with 79.9%. Additionally, the performance of the model is reported on each dataset after generating 8 paraphrases for each context (see Section 4.1). It is observed there is a consistent drop in performance on the paraphrased contexts compared to the original. This is expected as the nature of the machine generated paraphrased contexts do not necessarily align with the type of contexts observed in the original dataset. Table 3 further investigates the influence of the different elements: specifically the context influence compared to the question influence (note the question includes the options - see Section 3.1). The context of an MCRC question plays an important role in the output with influences up to 45% for MCTest.

The complexity of an MCRC question is described by the shape of the output distribution. A sharp distribution about the correct answer is indicative of an easy question while a flatter distribution over all the answer options indicates a harder question. Therefore, the strong influence of the context demonstrated in Table 3 emphasises that the context (alongside the specific posed question) is important in controlling the complexity of a question. To further verify the influence of the context, Figure 3 plots the complexity score output by the data complexity classifier. In particular, the distribution is shown for the complexity scores on the different subsets from RACE++ (for CMCQRD see Appendix Figure 10) of different complexity levels. Note, the normalized ranks of the complexity scores is plotted where the global rank is found for a given complexity score and divided by the total number of examples. The distributions are shown for the standard system (context, question and options), context-question system (context and question) and the context-only system. The context is clearly sufficient to determine the complexity levels of MC questions for these datasets, empirically supporting the importance of the context.

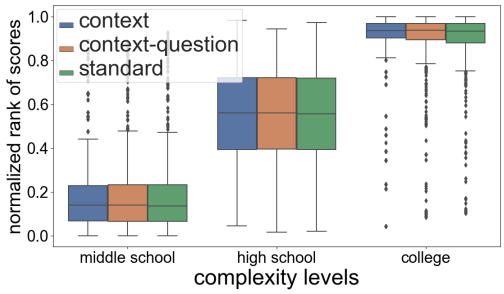


Figure 3: Normalized ranks (rank / total examples) of complexity scores for each complexity level using 3 evaluators: context, context-question and standard. See Appendix G.1.1 for the performance.

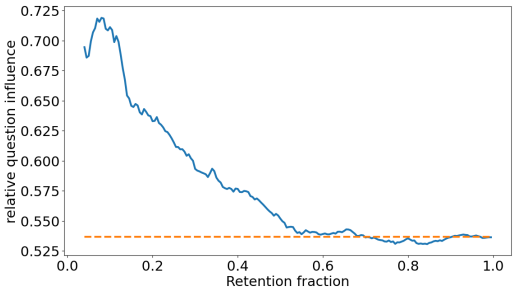


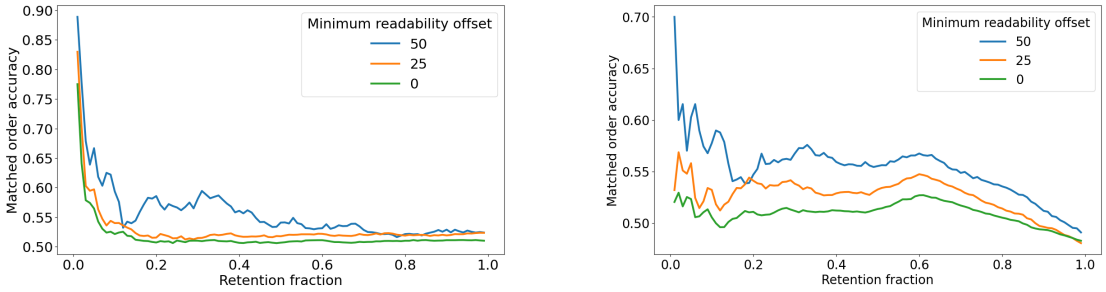
Figure 4: The relative question influence changes with the subset chosen by the rank of context complexity.

For the context, Table 3 further reports the influence for the semantic and linguistic components. For all 3 datasets, the semantic meaning of a context has a greater influence on the final output distribution but the specific linguistic realization also influences the output. Specifically, the relative semantic influence is greatest for MCTest and lowest

[3]Available at https://huggingface.co/wrmurray/roberta-base-finetuned-imdb

dataset	accuracy		influence		
	original	para	total	semantic	linguistic
IMDB	94.8	93.4	0.472	0.444 (94.0%)	0.028 (6.0%)
Yelp	94.3	93.9	0.472	0.445 (94.2%)	0.027 (5.8%)
Amazon	91.0	89.5	0.361	0.325 (90.0%)	0.036 (10.0%)

Table 4: Decomposition of input influence for different models in various datasets for sentiment classification.



(a) MCRC (b) SC

Figure 5: Entropy filtered pairwise agreement in paraphrase readability and true class probability ordering with various minimum readability gaps.

for CMCQRD. This is supported by Table 2 where the calculated semantic diversity is the lowest for MCTest with similar linguistic diversities across the datasets. Additionally, Table 3 suggests a relationship between the difficulty of a dataset (indicated by the accuracy of the model) and the relative question influence. The relative question influence increases with more challenging datasets. To further explore this observation, Figure 4 determines whether the question influence is directly linked to the complexity of a question. The data complexity classifier (QC system) is used to rank all the contexts according to their complexity. Then retaining a certain fraction of the most complex contexts, the relative question influence is plotted. The plot is for all the datasets combined (for RACE++ see Appendix Figure 11). Compared to the random ordering, it is clear that retaining the most complex contexts leads to larger question influence scores. The increase in question influence with more challenging contexts supports the trend from Table 3. Therefore, a more challenging context allows a greater variation in question difficulties, leading to a greater question influence on the output.

6.2 Sentiment classification

For the SC task there is only a single input. Therefore, we focus on exploring the relative contributions of the semantic and linguistic components of the input. Intuitively, the sentiment of a passage of text should be determined solely by its semantic content. However, Table 4 shows that for 3 popular SC datasets, the linguistic realization does influence the output. In particular, the relative linguistic component for Amazon hits 10% of the total. It appears the semantic component is more dominant for IMDb and Yelp compared to Amazon. One possible reason is that the length of texts is shorter for Amazon compared to the other datasets (see Table 2). Hence, longer texts have a greater opportunity to reinforce sentiment being expressed, which makes it more robust to different linguistic realizations. Appendix G.2 further explores this hypothesis by considering additional SC datasets.

6.3 Impact of linguistic realization

Section 6.1 demonstrated that the linguistic realization of the context in MCRC has measurable influence on the output distribution over the options. Here, we investigate the correlation between the readability of a given paraphrase of the original text and the output probability of the true class, termed true class probability (TCP). An entropy filter is applied to remove examples for which the entropy of the output distribution is too high, as high entropy examples suggest a random guess and hence challenging to ascertain whether a correlation exists. Figure 5a sweeps the fraction of examples retained according to the entropy filter and plots the fraction of the remaining examples for

which the ordering of TCP scores for every pair of paraphrases matches the ordering of their real readability scores. The plots are indicated for minimum readability gaps of 0, 25 and 50 for the pairs of paraphrases. It is observed that the readability of a paraphrase corresponds to the returned TCP, with a stronger correlation when the minimum readability gap between the pairs of paraphrases is higher. A similar process is applied for SC in Figure 5b to determine the relationship between the readability of the linguistic realization of the input and the TCP. Like MCRC there is a positive correlation between the readability level and the TCP, with a more pronounced relationship by constraining the pairs of paraphrases to have a larger readability gap.

7 Conclusions

This work describes an information-theoretic framework for text classification tasks. The framework determines the influence of each input element on the output. Additionally, each element is partitioned into its semantic and linguistic components. MCRC and SC are considered as case study tasks for analysis. For MCRC, it is found that both the context and question elements play influential roles on the output distribution. It is further established that selection of more challenging contexts permits greater variation (in terms of complexity) of questions on the context. Simpler contexts limit the range of the complexity to only easy questions. Hence, content creators need to carefully consider the choice of the context when designing MC questions to cater to a range of difficulty levels. In SC, the linguistic realization of the input has a measurable impact on the output. Hence, the text wording cannot be neglected when deducing the sentiment. For both tasks, higher the readability of a specific linguistic realization, greater the probability of the true class in the output distribution.

8 Acknowledgements

This research is partially funded by the EPSRC (The Engineering and Physical Sciences Research Council) Doctoral Training Partnership (DTP) PhD studentship and supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

9 Limitations

This work has several assumptions that must be stated. For the multiple-choice reading comprehension analysis, the question influence is based on real questions generated by humans on the original context. However, there is the possibility that the set of questions on a given context are not generated independently but instead the question creator has curated the question set together. Additionally, it is assumed that the paraphrasing of texts only changes the linguistic realization. However, it is likely that it also has an impact on the semantic content to an extent, which is reflected in the linguistic component influence on the output. We do quantify the appropriateness of the paraphrasing system in the Appendix.

10 Ethics statement

There are no ethical concerns with this work.

References

Ibrahim AlShourbaji, Na Helian, Yi Sun, Abdelazim G Hussien, Laith Abualigah, and Bushra Elnaim. 2023. An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. *Scientific Reports*, 13(1):14441.

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.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Luca Benedetto. 2023. A quantitative study of nlp approaches to question difficulty estimation. *arXiv preprint arXiv:2305.10236*.

Oscar Chew, Kuan-Hao Huang, Kai-Wei Chang, and Hsuan-Tien Lin. 2023. Understanding and mitigating spurious correlations in text classification. *arXiv preprint arXiv:2305.13654*.

KR1442 Chowdhary and KR Chowdhary. 2020. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Vinícius G Costa and Carlos E Pedreira. 2023. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, Dayheon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-agnostic question generation for education. *arXiv preprint arXiv:2203.08685*.

Jianqing Fan and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2023. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. *arXiv preprint arXiv:2309.12551*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Yifan Gao, Lidong Bing, Wang Chen, Michael R Lyu, and Irwin King. 2018. Difficulty controllable generation of reading comprehension questions. *arXiv preprint arXiv:1807.03586*.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6423–6430.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Ark, Vander. 2012. The hewlett foundation: Automated essay scoring.

Chao Huang, Diptesh Das, and Koji Tsuda. 2023. Feature importance measurement based on decision tree sampling. *arXiv preprint arXiv:2307.13333*.

Yejin Hwang and Jongwoo Song. 2023. Recent deep learning methods for tabular data. *Communications for Statistical Applications and Methods*, 30(2):215–226.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and E. Hovy. 2017a. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017b. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Yichan Liang, Jianheng Li, and Jian Yin. 2019a. 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.

Yichan Liang, Jianheng Li, and Jian Yin. 2019b. A new multi-choice reading comprehension dataset for curriculum learning. In *Asian Conference on Machine Learning*, pages 742–757. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2023a. Zero-shot nlg evaluation through pairware comparisons with llms. *arXiv preprint arXiv:2307.07889*.

Adian Liusie, Vatsal Raina, and Mark Gales. 2023b. " world knowledge" in multiple choice reading comprehension. In *The Sixth Fact Extraction and VERification Workshop*, page 49.

Adian Liusie, Vatsal Raina, Andrew Mullooly, Kate Knill, and Mark J. F. Gales. 2023c. Analysis of the cambridge multiple-choice questions reading dataset with a focus on candidate response distribution.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 78–84.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. *arXiv preprint arXiv:2301.12307*.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J.F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. 2023. The Cambridge Multiple-Choice Questions Reading Dataset. Cambridge University Press and Assessment.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE.

Vatsal Raina and Mark Gales. 2022a. Answer uncertainty and unanswerability in multiple-choice machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1020–1034, Dublin, Ireland. Association for Computational Linguistics.

Vatsal Raina and Mark Gales. 2022b. Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*.

Vatsal Raina, Nora Kassner, Kashyap Popat, Patrick Lewis, Nicola Cancedda, and Louis Martin. 2023a. ERATE: Efficient retrieval augmented text embeddings. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 11–18, Dubrovnik, Croatia. Association for Computational Linguistics.

Vatsal Raina, Adian Liusie, and Mark Gales. 2023b. Analyzing multiple-choice reading and listening comprehension tests. *arXiv preprint arXiv:2307.01076*.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013a. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013b. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel R Bowman. 2022. What makes reading comprehension questions difficult? *arXiv preprint arXiv:2203.06342*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. *arXiv preprint arXiv:2310.14542*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

Biao Xu, Yao Wang, Xiuwu Liao, and Kaidong Wang. 2023. Efficient fraud detection using deep boosting decision trees. *Decision Support Systems*, page 114037.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Additional Derivations

In Equations 13 and 14, we demonstrate the decomposition of influence for the MCRC task, following the framework established in Equations 2 and 3. Equations 5 through 8 detail the theoretical computation of various mutual information metrics. The rationale for Equation 6 is as follows:

$$
\begin{aligned}
&I\left(Y; X_j\right) \\
&= \mathcal{H}\left(P(y)\right) - \mathbb{E}_{P(x_j)}\left[\mathcal{H}(P(y|x_j))\right] \\
&= \mathcal{H}\left(\mathbb{E}_{P(x)}[P(y|x)]\right) - \mathbb{E}_{P(x_j)}\left[\mathcal{H}(P(y|x_j))\right]
\end{aligned}
$$

Similarly, Equations 13 and 14 can be written as:

$$
\begin{aligned}
\mathcal{I}\left(Y; C, Q\right) = \mathcal{H}\left(\mathbb{E}_{P(c,q)}[P(y|c,q)]\right) \\
- \mathbb{E}_{P(c,q)}\left[\mathcal{H}\left(P\left(y|c,q\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}\left(Y; C\right) = \mathcal{H}\left(\mathbb{E}_{P(c,q)}[P(y|c,q)]\right) \\
- \mathbb{E}_{P(c)}\left[\mathcal{H}\left(P\left(y|c\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}\left(Y; Q|C\right) = \mathbb{E}_{P(c)}\mathcal{H}\left([P(y|c)]\right) \\
- \mathbb{E}_{P(c,q)}\left[\mathcal{H}\left(P\left(y|c,q\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}\left(Y; C^s\right) = \mathcal{H}\left(\mathbb{E}_{P(c,q)}[P(y|c,q)]\right) \\
- \mathbb{E}_{P\left(c^{(s)}\right)}\left[\mathcal{H}\left(P\left(y|c^{(s)}\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}\left(Y; C|C^s\right) = \mathbb{E}_{P\left(c^{(s)}\right)}\left[\mathcal{H}\left(P\left(y|c^{(s)}\right)\right)\right] \\
- \mathbb{E}_{P(c)}\left[\mathcal{H}\left(P\left(y|c\right)\right)\right]
\end{aligned}
$$

There are 4 terms to be calculated:

$$
\mathbb{E}_{P(c,q)}[\mathcal{H}(P(y|c,q))], \quad \mathcal{H}\left(\mathbb{E}_{P(c,q)}[P(y|c,q)]\right)
$$

$$
\mathbb{E}_{P(c)}[\mathcal{H}(P(y|c))], \quad \mathbb{E}_{P(c^{(s)})}[\mathcal{H}\left([P(y|c^{(s)})]\right)
$$

They are approximated in Equations 16 to 19. Take Equation 16 as an example:

$$
\begin{aligned}
&\mathbb{E}_{P(c,q)}[P(y|c,q)] = \int\int P(y|c,q)P(c,q)dqdc \\
&= \int\int P(y|c,q)P(q|c)P(c)dqdc \\
&= \int\int\int P(y|c,q)P(q|c)P(c|c^s)P(c^s)dqdcdc^s
\end{aligned}
$$

The integral is intractable. Thus we use Monte Carlo as an approximation. For the innermost integral:

$$
\int P(y|c,q)P(q|c)dq \approx \frac{1}{|Q|}\sum_{q\sim P(q|c)} P(y|c,q)
$$

Similarly, with Monte Carlo, the whole equation can be approximated as:

$$
\frac{1}{|C^{(s)}|}\sum_{c^{(s)}}\frac{1}{|C|}\sum_{c\sim P(c|c^{(s)})}\frac{1}{|Q|}\sum_{q\sim P(q|c)} P(y|c,q)
$$

In practice, given a dataset, we have n_s data points (the i^{th} data point contains a context realization $\tilde{r}_i$ which is an example of context element c^i or a realization of $c_i^{(s)}$, and a number of $|\tilde{Q}_i|$ question-option pairs). Thus we have $|C^{(s)}| = n_s$. As mentioned in section 3.1.2, we use ChatGPT to generate paraphrases r_i from the observed context $\tilde{r}_i$ as different linguistic realizations given the same semantic meaning. (performance of paraphrasing generation system is in Appendix E). Thus we have $|C| = |R^i|$ and $c \sim P(c|c^{(s)}) \rightarrow r \in R^{(i)}$. The observed questions $\tilde{Q}^{(i)}$ can be seen as $q \sim P(q|c)$ directly. Thus we have $q \sim P(q|c) \rightarrow \tilde{q} \in \tilde{Q}$ with $|Q| = |\tilde{Q}^{(i)}|$. Overall, we get:

$$
\frac{1}{n_s}\sum_{i=1}^{n_s}\frac{1}{|R^{(i)}|}\sum_{r\in R^{(i)}}\frac{1}{|\tilde{Q}^{(i)}|}\sum_{\tilde{q}\in\tilde{Q}} P(y|r,\tilde{q})
$$

Or

$$
\frac{1}{n_s}\sum_{i=1}^{n_s}\frac{1}{|R^{(i)}||\tilde{Q}^{(i)}|}\sum_{r\in R^{(i)},\tilde{q}\in\tilde{Q}} P(y|r,\tilde{q})
$$

Similar procedure works for equations 17 to 23.

B Extended related work

In multiple-choice reading comprehension, the influence of each element on the final output distribution is directly linked to the complexity of a multiple-choice question. More complex multiple-choice questions can expect to have flat distributions over the answer options while easier questions are sharp around the correct answer. Numerous studies have looked at the factors that potentially influence complexity of context passages in relation with reading comprehension tasks. Sugawara et al. (2022) observed that the diversity of contexts in MC questions determines the diversity of questions possible conditioned on the contexts. In their work they found that variables such as passage source, length, or readability measures do not significantly affect the model performance. Further, Khashabi et al. (2018) introduced the role of the original source from which the contexts are extracted in shaping overall complexity. Through the augmentation of the dataset by diversifying the

corpus sources, they aimed to enhance the dataset quality.

Question complexity has repeatedly been discussed within prior literature, with varying definitions. Liang et al. (2019b) classified questions into distinct categories with complexity scores ranking from lowest to highest for word matching, paraphrasing, single-sentence reasoning, multi-sentence reasoning, and ambiguous questions. Gao et al. (2018) quantified complexity as the number of reasoning steps required to derive the answer. Similar definitions have been upheld by Yang et al. (2018) and Dua et al. (2019), who expanded the understanding of question complexity to encompass not only contextual comprehension but also factors such as the confidence of a pretrained question-answering model.

Distractor (incorrect options for multiple-choice questions) complexity has been less explored. Gao et al. (2019) determined distractor complexity based on the similarity between distractors and the ground-truth. Employing an n-gram overlap metric, Banerjee and Lavie (2005) introduced a method to assess distractor complexity. Dugan et al. (2022) further dissected distractor complexity, analyzing qualities such as relevance, interpretability and acceptability compared to human markers.

As not explored in previous MCRC complexity literature, in this work the information-theoretic approach is applied to characterize the influence of each element in a given multiple-choice reading comprehension dataset. Greater the influence of an element, greater the scope to control the complexity of the multiple-choice reading comprehension task.

In sentiment classification, it is expected the sentiment class should be dependent on the semantic meaning of the text rather than its linguistic realization. However, Liusie et al. (2022) find that *shortcut* systems that have access only to the stop-words in the original text are also able to identify the sentiment class. Hence, they find the stop words chosen in the text do influence the sentiment class, which we express as the specific linguistic realization. Chew et al. (2023) further aim to correct for the bias from spurious correlations. In this work, we explicitly quantify the influence of the semantic and linguistic components of the text.

C Additional Tasks

C.1 Grade classification

In the grade classification task, the system is input a prompt z and a response x and then is required to output a number in the range 1 to 6 with 6 denoting the highest degree of alignment between the given prompt and its corresponding response. The task is traditionally a regression-oriented but here we apply our information-theoretic classification framework by treating it as a 6-option classification task.

$$P(y) = \mathbb{E}_{P(z,x)} P(y|z,x) \tag{25}$$

Further, the semantic meaning of the response x can be seen as generated from $P(x|z)$. In the considered datasets, a prompt has a large-scale number of responses while the number of prompts are limited. Thus, here we focus only on the analysis of the influence of the response. The output equation can be rewritten as:

$$P(y) = \mathbb{E}_{P(z)} \mathbb{E}_{P(x^{(s)}|z)} \mathbb{E}_{P(x|x^{(s)})} P(y|x,z) \tag{26}$$

Therfore, we can calculate the influence from the semantic meaning and the linguistic realization of the responses:

$$\underbrace{\mathcal{I}(Y;X|Z)}_{\text{text}} = \underbrace{\mathcal{I}(Y;X^{(s)}|Z)}_{\text{semantic}} + \underbrace{\mathcal{I}(Y;X|X^{(s)},Z)}_{\text{linguistic}} \tag{27}$$

In practice, we collect the prompt set $\mathcal{Z}$ and a response set $\mathcal{X}$ for each prompt. For each response $x^{(i,j)}$ of the prompt $z^{(i)}$, we observe a realization $\tilde{r}^{(i,j)}$ and additionally generate a set of realizations $\mathcal{R}^{(i,j)}$ of the semantic meaning of the responses. The following approximations are made:

$$\mathbb{E}_{P(z)} \left[\mathcal{H} \left(\mathbb{E}_{P(x|z)} \left[P(y|x,z) \right] \right) \right] \approx \tag{28}$$

$$\frac{1}{|\mathcal{Z}|} \sum_{z^i \in Z} \mathcal{H}\Big(\frac{1}{n_{\text{s}}^i} \sum_{j=1}^{n_{\text{s}}^i} \frac{1}{|\mathcal{R}^{(i,j)}|} \sum_{r \in \mathcal{R}^{(i,j)}} P(y|r,z^i) \Big)$$

$$\mathbb{E}_{P(x^{(s)},z)} \Big[\mathcal{H}(P(y|x^{(s)},z)) \Big] \approx \tag{29}$$

$$\frac{1}{|\mathcal{Z}|} \sum_{z^i \in Z} \frac{1}{n_{\text{s}}^i} \sum_{j=1}^{n_{\text{s}}^i} \mathcal{H}\Big(\frac{1}{|\mathcal{R}^{(i,j)}|} \sum_{r \in \mathcal{R}^{(i,j)}} P(y|r,z^i) \Big)$$

$$\mathbb{E}_{P(x,z)} \left[\mathcal{H}(P(y|x,z)) \right] \approx \tag{30}$$

$$\frac{1}{|\mathcal{Z}|} \sum_{z^i \in Z} \frac{1}{n_{\text{s}}^i} \sum_{j=1}^{n_{\text{s}}^i} \frac{1}{|\mathcal{R}^{(i,j)}|} \sum_{r \in \mathcal{R}^{(i,j)}} \mathcal{H}(P(y|r,z^i))$$

where n_s^i is the number of responses for prompt $z^{(i)}$.

D Systems

D.1 Grade classification

The architecture of the model used for grade classification is based on the decoder-only transformer architecture with Llama-2 as described in Section 4.2. Hence, the prompt and the response are concatenated together at the input to the model, which is trained to return a probability distribution over the 6 grade classes.

D.2 Data complexity classification

The system is required to output a probability distribution among options: easy, medium and hard. The architecture is presented in Figure 6. The complexity score S_c is calculated as:

$$S_c = 0 * P_{easy} + 0.5 * P_{medium} + 1 * P_{hard} \quad (31)$$

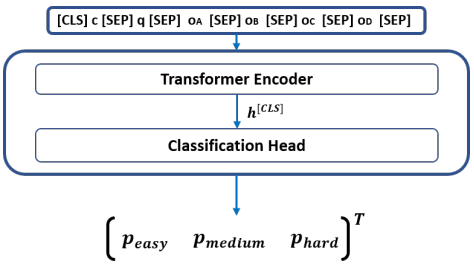


Figure 6: Architecture for MC question complexity classifier with context, c, question, q and options, $\{o\}$.

D.3 Calibration

The trained models were calibrated post-hoc using single parameter temperature annealing (Guo et al., 2017). It is necessary to calibrate the models for the absolute information-theoretic measures to be meaningful. Uncalibrated, model probabilities are determined by applying the softmax to the output logit scores s_i:

$$P(y = k; \boldsymbol{\theta}) \propto \exp(s_k) \quad (32)$$

where k denotes a possible output class for a prediction y. Temperature annealing 'softens' the output probability distribution by dividing all logits by a single parameter T prior to the softmax.

$$P_{CAL}(y = k; \boldsymbol{\theta}) \propto \exp(s_k/T) \quad (33)$$

As the parameter T does not alter the relative rankings of the logits, the model's prediction will be unchanged and so temperature scaling does not affect the model's accuracy. The parameter T is chosen such that the accuracy of the system is equal to the mean of the maximum probability (as is expected for a calibrated system).

E Experiments

E.1 Data

Table 5 provides a breakdown of the RACE++ dataset, which is divided into three subsets: RACE-M, RACE-H, and RACE-C. These subsets correspond to English exam materials from Chinese middle schools (RACE-M), high schools (RACE-H), and colleges (RACE-C) respectively. Table 5 displays key statistics for each of these subsets including the number of contexts, the average number of questions for one context, and the semantic and linguistic diversity. Note, for all MCRC datasets, the options are re-ordered such that the true class is the first option.

SST-2 (Socher et al., 2013) and TweetEval (Barbieri et al., 2020) are considered as additional datasets for sentiment classification, which were not presented in the main text. SST-2 (Stanford Sentiment Treebank) corpus consists of movie reviews provided by Pang and Lee (2005) which are classified as either positive or negative. TweetEval consists of seven heterogeneous tasks based on tweets from Twitter. Here, the focus is on the tweet-emotion task where each tweet is classified as joy, sadness, optimism or anger. These two datasets are included here as they have shorter inputs texts than IMDb, Yelp and Amazon.

Hewlett foundation (Hamner et al., 2012), a competition requiring automated grading of student-written essays, is included as the dataset for the grade classification task. In total it has 8 subgroups and each subgroup has 1 prompt with several responses. Here we only choose the first 2 subgroups and turn them into a 6 option classification task with 6 denoting the highest degree of alignment between the given prompt and its corresponding response. 3,583 responses are sampled for the training split while 500 are selected randomly as the test split. From Table 5, it is clear compared to other dataset, the responses in Hewlett are mutually semantically closer in meaning compared to other datasets as they are all responses to just 2 prompts.

Table 5 also supports that our paraphrasing system is sensible by showing the linguistic diversity is much lower than the semantic diversity among all responses. A more detailed quality verification

process of our paraphrasing system is shown in Appendix F.

E.2 Models

For the multiple choice reading comprehension tasks, three models are used: Llama2, RoBERTa and Longformer.

Pretrained Llama-2 (7 billion parameters) is finetuned specifically on the train split of RACE++ with hyperparameter tuning on the validation split. However, it is not computationally feasible to train all the model parameters of Llama-2. Therefore, parameter efficient finetuning is used with quantized low rank adapters (QLoRA) (Dettmers et al., 2023). The final training parameters finetune the model with a learning rate of 1e-4, batch size of 4, lora rank of 8 with lora α = 16 and dropout 0.1. The model is trained for 1 epoch taking 7 hours on an NVIDIA A100 machine. For the main paper results, the Llama-2 model is selected due to its best accuracy.

RoBERTa (82 million parameters)[4] is finetuned on the train split of the RACE dataset (RACE-M and RACE-H). The details of the specific Longformer (336 million parameters)[5] are detailed in Manakul et al. (2023). For the main paper results, the Llama-2 model is selected due to its best accuracy.

For the data complexity classification system, pretrained ELECTRA-base (110 million parameters) is finetuned on the RACE++ train split with the complexity class (easy, medium or hard) as the label. The model is trained using the AdamW optimizer, a batch size of 3, learning rate of 2e-5, max number of epochs of 3 with all inputs truncated to 512 tokens. An ensemble of 3 models is trained. Training for each model takes 3 hours on an NVIDIA V100 graphical processing unit.

For the sentiment classification task, we used the models from RoBERTa and distilBERT(82 million parameters) (Sanh et al., 2019) family and the they are finetuned on various datasets as explained in the following. The train split of IMDb is used to finetune RoBERTa[6] and BERT[7]. Both of these models are applied on the test sets for IMDb, Yelp and Amazon. The models we used for SST-2 and TweetEval are finetuned n their corresponding training split, namely RoBERTa-SST2[8], distilBERT-SST2[9], RoBERTa-Tweet[10], distilBERT-Tweet[11].

For the grade classification task, we finetuned Llama-2 on the training split of the Hewlett dataset using QLoRA. The chosen training parameters finetune the model with a learning rate of 1e-4, batch size of 4, lora rank of 8 with lora α = 16 and dropout 0.1. The model is trained for 1 epoch taking 30 minutes on an NVIDIA A100 machine. The sentiment classification datasets do not state licenses.

E.3 Question filter

Based on manual observation, the RACE++ dataset has some questions that are generated from the linguistic contents of the contexts rather than from the semantic contents. As explained in Section 3.1, these questions will invalidate the theoretical assumption when calculating the influence of each component because the question would be unanswerable for a generated paraphrase that does not maintain the same linguistic information. Namely, these linguistic questions are often related to their positions in the context and are always correlated with certain key words such as 'in paragraph 2'. Thus, we apply a word-matching question filter to filter out all such examples, ensuring that only relevant and contextually coherent questions are retained for further processing. We specifically filter out all questions containing the following phrases: '{number} + word/sentence/paragraph + {number} + refer to/mean'.

In total, for the RACE++ dataset, approximately 6.2% questions are found to be generated from the linguistic content of the context and thus filtered out. The effects of the filter on the element influence analysis using Llama-2 is shown in Table 6. It is clear the measured question influence drops as expected by 1.6%.

[4]Available at https://huggingface.co/LIAMF-USP/roberta-large-finetuned-race

[5]Available at https://huggingface.co/potsawee/longformer-large-4096-answering-race

[6]Available at: https://huggingface.co/wrmurray/roberta-base-finetuned-imdb

[7]Available at : https://huggingface.co/lvwerra/distilbert-imdb

[8]Available at: rasyosef/roberta-base-finetuned-sst2

[9]Available at: https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

[10]Available at: cardiffnlp/twitter-roberta-base-dec2021-emotion

[11]Available at: https://huggingface.co/philschmid/DistilBERT-tweet-eval-emotion

		# examples	# options	# words	# questions	semantic diversity	linguistic diversity
MCRC	RACE-M	362	4	200	4	$0.096_{\pm 0.017}$	$0.017_{\pm 0.007}$
	RACE-H	510	4	308	3.3	$0.100_{\pm 0.012}$	$0.016_{\pm 0.006}$
	RACE-C	135	4	375	5.2	$0.097_{\pm 0.011}$	$0.018_{\pm 0.006}$
SC	SST-2	500	2	20	-	$0.145_{\pm 0.021}$	$0.087_{\pm 0.034}$
	TweetEval	500	4	17	-	$0.149_{\pm 0.022}$	$0.081_{\pm 0.033}$
GC	Hewlett	500	6	383	2	$0.063_{\pm 0.017}$	$0.021_{\pm 0.009}$

Table 5: Statistics for breakdown of additional test datasets including RACE-M, RACE-H, RACE-C for RACE++ in multiple-choice reading comprehension (MCRC); SST-2, TweetEval in sentiment classification (SC) and Hewlett in grade classification (GC).

filter	accuracy		influence				
	original	para	total	question	context	context-semantic	context-linguistic
No	84.2	81.5	0.284	0.164 (57.7%)	0.120 (42.3%)	0.135 (81.4%)	0.031 (18.6%)
Yes	86.0	82.9	0.298	0.161 (56.1%)	0.131 (43.9%)	0.108 (82.5%)	0.023 (17.5%)

Table 6: The effect of the question filter on element influence for Llama-2 on the RACE++ test set.

F Paraphrasing

The readability level is measured by Flesch reading-ease (Flesch, 1948) score (FRES) where higher scores indicate material that is easier to read while lower scores are reflective of more challenging texts.

$$\texttt{FRES} = 206.835 - 1.015\left(\frac{n_w}{n_{se}}\right) - 84.6\left(\frac{n_{sy}}{n_w}\right)$$

n_w is the total number of words, n_{se} is the total number of sentences, n_{sy} the total number of syllables.

We grouped the original texts into eight different readability levels: 5, 20, 40, 55, 65, 75, 85, and 95 for the reading comprehension and grade classification tasks and used the final 7 groups for the sentiment classification task as there were no texts in sentiment classification that fell into the most challenging category of 0-10 on the readability scale. The specific prompts for each difficulty level we used are shown in Table 7. Here we also present the quality of our paraphrase generation process. Figure 7 displays the average readability score of the paraphrased text for each combination of original and target readability levels. From the heatmap, we can see that while the readability of the paraphrased text is influenced by the readability of the original text, the paraphrases still fall within an acceptable range of readability. We also report the averaged BertScore F1 (Zhang et al., 2019) and Word Error Rate (WER) (Och, 2003) to ensure the quality of our paraphrasing system as shown in Figure 8 and Figure 9. An ideal paraphrasing system should expect high semantic similarity with high BERTScore and low linguistic similarity with high WER.

G Additional results

Here we present the results from some additional experiments that act as a supplement to the main paper.

G.1 Reading comprehension

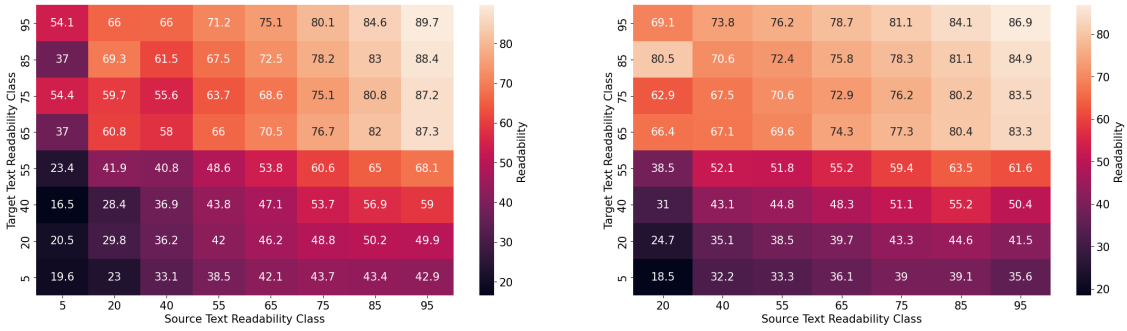
G.1.1 Data complexity classifier

In Section 4.3, we used the data complexity classifier with the data context as the input. Here we assess its quality by testing its performances in-domain (with an ensemble of 3 models) on the RACE++ test set. We additionally compare the performance on the standard input with other possible combinations of the input, as shown in Table 8. As well as accuracy, macro F1 is reported to account for the imbalance in the complexity level classes. The results for the mode class indicate the baseline performance when the mode class is selected for every example in the test set. All systems significantly outperform the baseline. Inputting the context alone is sufficient to get an accuracy close to the full input and when extra information is inputted, the gain is marginal. Hence, compared with question and options, the context carries a substantial proportion of the information to determine the complexity of a question.

In Figure 10, the data complexity classifier model shows strong generalizability by clearly classifying different subsets of CMCQRD dataset,

Target	Level (US)	Prompt
5	Professional	Paraphrase this document for a professional. It should be extremely difficult to read and best understood by university graduates.
20	College graduate	Paraphrase this document for college graduate level (US). It should be very difficult to read and best understood by university graduates.
40	College	Paraphrase this document for college level (US). It should be difficult to read.
55	10-12th grade	Paraphrase this document for 10th-12th grade school level (US). It should be fairly difficult to read.
65	8-9th grade	Paraphrase this document for 8th/9th grade school level (US). It should be plain English and easily understood by 13- to 15-year-old students.
75	7th grade	Paraphrase this document for 7th grade school level (US). It should be fairly easy to read.
85	6th grade	Paraphrase this document for 6th grade school level (US). It should be easy to read a
95	5th grade	Paraphrase this document for 5th grade school level (US). It should be very easy to read and easily understood by an average 11-year old student.

Table 7: Prompts to generate paraphrases with different target readability (using FRES).



(a) Reading Comprehension

(b) Sentiment Classification

Figure 7: Averaged measured readability.

input format	accuracy single	accuracy ens	F1 single	F1 ens
mode class	61.6	–	25.4	–
standard	$84.7_{\pm 0.5}$	87.2	$81.7_{\pm 1.1}$	83.7
context	$84.9_{\pm 0.3}$	85.1	$81.8_{\pm 0.8}$	81.7
context-question	$84.7_{\pm 0.7}$	86.0	$81.8_{\pm 0.6}$	82.2
question-option	$70.2_{\pm 0.5}$	71.3	$67.3_{\pm 0.7}$	68.2

Table 8: Accuracy of data complexity evaluators on the RACE++ test set.

which differ a lot from the model's training dataset. The plot also supports the context is a sufficient input to determine the different complexity levels.

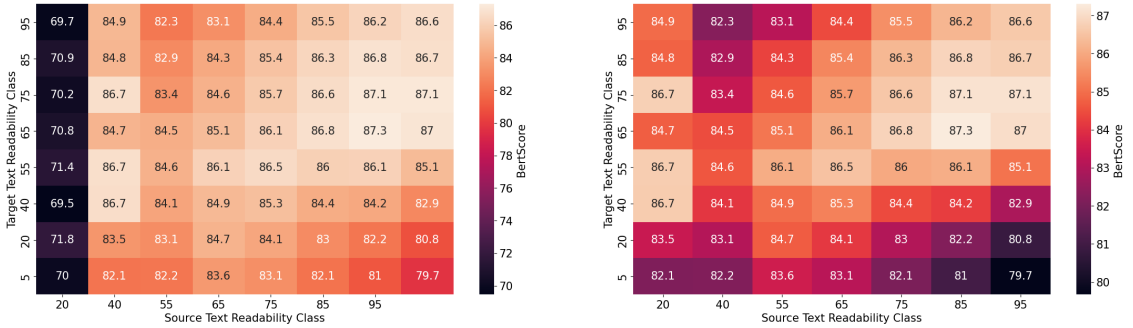
G.1.2 Additional models

In Section 6, we analyse the influence from different elements and components on the output for two specific models: Llama-2 for the multiple choice reading comprehension task, Roberta for the sentiment classification task. Here we show the influence terms calculated are not model-specific by showing the consistency of the element influences on the same datasets but evaluated by different models: Llama-2, Roberta and Longformer as in Table 9.

G.1.3 Further analysis

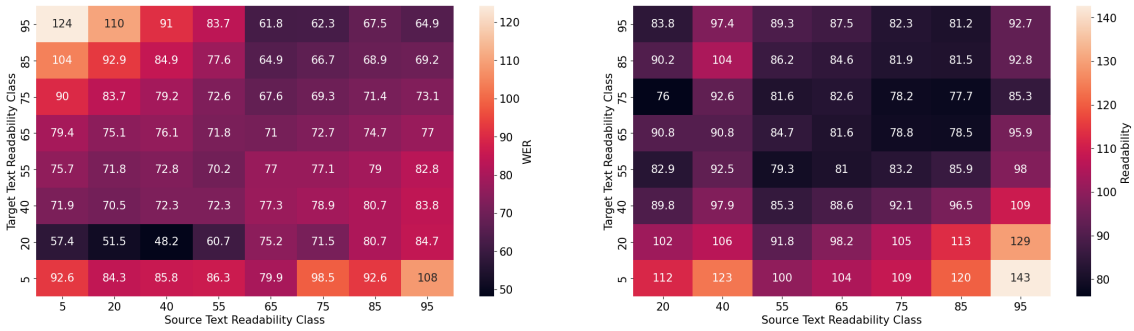
In Figure 4, we show a strong positive correlation between the question influence and the context complexity when we consider all three dataset together. Here we show the rule still holds for data points inside a single dataset in Figure 11 where the trend in RACE++ dataset is pretty similar to the all three datasets together.

We also explore other potential factors influencing the relative question influences. From Table 2, there are two marked differences between the datasets: the number of words (length) and the number of questions. To find their influence in the relative question influence score, Figure 12a and



(a) Reading Comprehension (b) Sentiment Classification

Figure 8: Averaged BERTScore F1.



(a) Reading Comprehension (b) Sentiment Classification

Figure 9: Averaged Word Error Rate.

Figure 12b show the relative question influence of the subset chosen from the contexts ranked by their number of words or the number of questions of the corresponding context. A strong positive trend between the relative question influence scores and the context length is observed as expected: a longer context naturally has a larger question generation capacity. As shown in Figure 12b, the number of questions does not have a direct impact, indicating the results are not affected by the specific number of questions per context.

G.1.4 Question Generation

In Section 3.1 we investigate the influence of different element of model input and focuses on human-generated questions only. Here we investigate the influence from the LLM-generated questions.

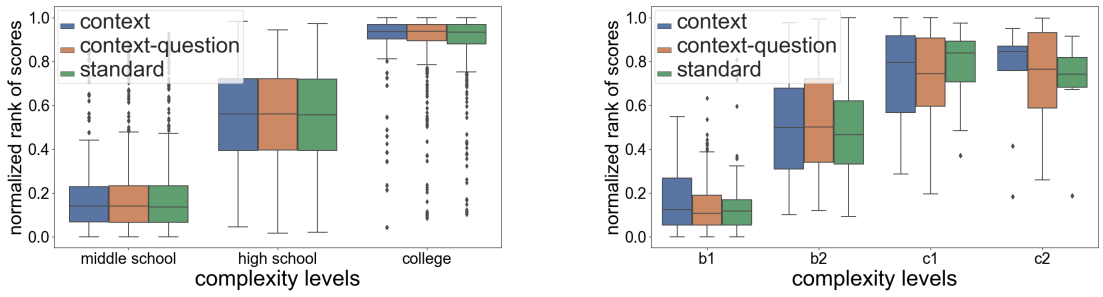
To be more specific, for each context, using GPT-4, we generate 4 questions with prompt:

"Given the context: {context}, please generate four multiple choice questions with options where the first option is the correct answer and the other three are distractors. The questions should be of varying difficulty levels: low, middle, high, and very high. Please output the questions in the format of a dictionary with the keys: 'easy', 'middle', 'high', and 'very high'. Each key should map to a dictionary representing a question, with the fields 'question', 'options', and 'answer' indicating the correct answer."

However, as indicated in (Sun et al., 2023), LLMs struggle at tasks with hard constraints, Table 10 shows the generated questions are relatively easy and have a lower influence on model output.

G.1.5 Ordering

For humans taking multiple-choice tests, the role of the context compared to the question may be influenced by the ordering in which they read each of these elements. Similarly, a reading comprehension system may be susceptible to the ordering of the context and the question. Here we compare the influence of the ordering by reversing the standard context followed by question at the input to the question followed by the context . Table 11 demonstrates that the ordering for the automated



(a) RACE++ (b) CMCQRD

Figure 10: Normalized ranks (rank / total examples) of complexity scores for each complexity level using three complexity evaluators: context, context-question and standard.

dataset	model	accuracy		influence				
		orig	para	total	question	context	c-semantic	c-linguistic
MCTest	RoBerta	95.3	93.0	0.254	0.140 (55.2%)	0.114 (44.8%)	0.076 (67.0%)	0.037 (33.0%)
	Longformer	98.3	91.3	0.285	0.152 (53.5%)	0.133 (46.5%)	0.068 (70.6%)	0.028 (29.4%)
	Llama2	92.5	85.8	0.212	0.116 (54.7%)	0.096 (45.3%)	0.068 (70.6%)	0.028 (29.4%)
RACE++	RoBerta	84.2	81.5	0.379	0.213 (56.3%)	0.166 (43.7%)	0.135 (81.4%)	0.031 (18.6%)
	Longformer	81.6	79.3	0.390	0.228 (58.6%)	0.162 (41.1%)	0.135 (83.2%)	0.027 (16.8%)
	Llama2	86.0	82.9	0.298	0.161 (56.1%)	0.131 (43.9%)	0.108 (82.5%)	0.023 (17.5%)
CMCQRD	Roberta	73.5	69.4	0.383	0.287 (74.9%)	0.096 (25.1%)	0.074 (77.5%)	0.022 (22.4%)
	Longformer	71.9	69.8	0.467	0.326 (69.9%)	0.141 (30.1%)	0.114 (81.0%)	0.027 (19.0%)
	Llama2	79.9	69.4	0.290	0.211 (72.7%)	0.079 (27.3%)	0.067 (83.8%)	0.012 (16.2%)

Table 9: Decomposition of total input influence for different models in various multiple-choice reading comprehension datasets. With c-semantic is context semantic, c-linguistic is context-linguistic

system does not lead to differing influences on each element. The results here are provided for the fine-tuned Llama-2 model from Section 5.2.

G.2 Sentiment classification

For the sentiment classification task, to show the consistency of the element influence among different models, Table 12 presents additional results using the BERT model as a comparison against the RoBERTa model.

It can be observed also that for shorter input text datasets, such as SST-2 and TweetEval, the linguistic component is more significant, approaching 20% of the total.

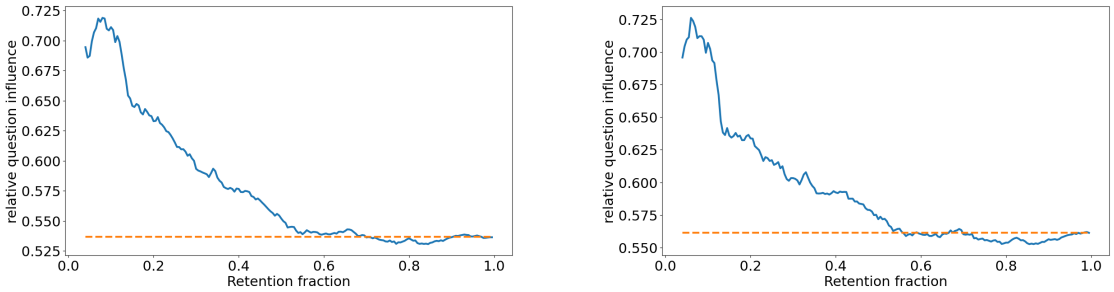
G.3 Grade classification

The influence to the model output from the response, its semantic component and linguistic component are respectively shown in Table 13. We observe a large influence from the linguistic content of the responses which agrees with the 15% drop in model accuracy when the paraphrased dataset is used. Further, we measure the correlation between the readability and true class probability in Figure 13. It is clear that with a higher readability, the probability of selecting the true grade increases.

H Future work

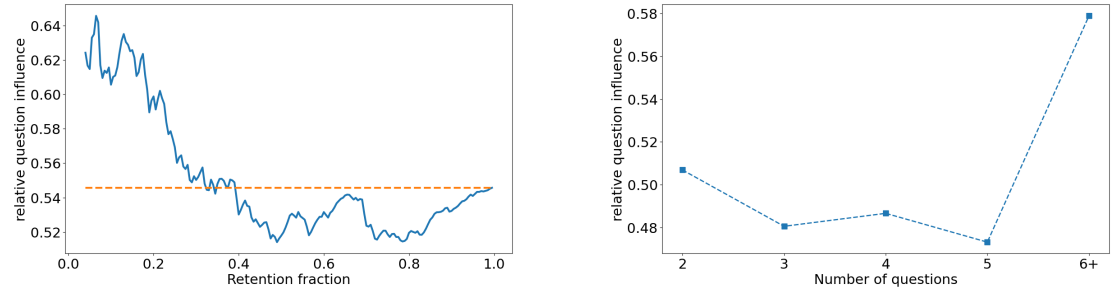
The analysis in this work has applied the framework to specifically NLP classification tasks. It would be interesting to extend the framework to both regression and sequence output tasks. For sequential outputs, there needs to be a methodology to convert the generated sequence to a single score such that its sensitivity can be measured to each input element.

The framework applied to textual data to explore the influence of semantic vs linguistic components can also be extended to image inputs. Here, we can perceive the semantic content as the object being described in the image while the linguistic realization is based on the recording equipment that controls aspects such as orientation, resolution (blurring), camera angle, e.t.c. Therefore, the proposed information-theoretic approach has potential applications across several modalities.



(a) Total (b) RACE++

Figure 11: The relative question influence changes with the subset chosen by the rank of context complexity in all three datasets (left) and in RACE++ only (right). [12] 0.2 in x-axis means we leave contexts with top 20% context complexity as the subset.



(a) Length (b) Number of questions

Figure 12: The relative question influence for a subset of contexts swept in order of length (left) or average number of questions per context (right) for all MCRC datasets with Llama-2.

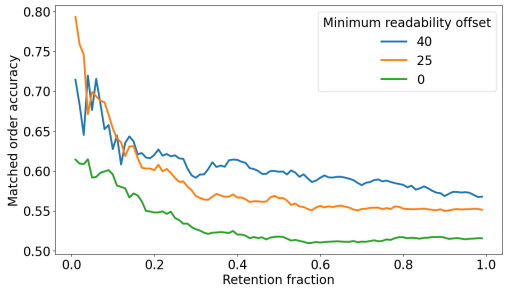


Figure 13: Entropy filtered pairwise agreement in paraphrase readability and true class probability ordering with various minimum readability gaps.

I Licenses

The RACE dataset is available for non-commercial research purposes only. Also for CMCQRD, the license[13] states the licensed dataset for non-commercial research and educational purposes only. MCTest is copyright free. There are no licenses stated for the remaining datasets.

[13]Available at: https://englishlanguageitutoring.com/datasets/cambridge-multiple-choice-questions-reading-dataset

Question Source	accuracy		influence				
	original	para	total	question	context	context-semantic	context-linguistic
human	84.2	81.5	0.284	0.164 (57.7%)	0.120 (42.3%)	0.135 (81.4%)	0.031 (18.6%)
automatic	96.1	93.7	0.266	0.121 (45.4%)	0.145 (54.6%)	0.101 (69.5%)	0.044 (30.5%)

Table 10: Human sources questions vs GPT4 generated questions for Llama-2 on the RACE++ test set.

dataset	direction	influence				
		total	question	context	context-semantic	context-linguistic
RACE++	Forward	0.304	0.171 (56.3%)	0.133 (43.7%)	0.109 (82.1%)	0.024 (17.9%)
	Reverse	0.305	0.172 (56.7%)	0.133 (43.6%)	0.109 (82.3%)	0.024 (17.7%)
MCTest	Forward	0.212	0.116 (54.7%)	0.096 (45.3%)	0.068 (70.6%)	0.028 (29.4%)
	Reverse	0.229	0.129 (56.6%)	0.100 (43.4%)	0.067 (67.2%)	0.032 (32.3%)
CMCQRD	Forward	0.290	0.211 (72.7%)	0.079 (27.3%)	0.067 (83.8%)	0.012 (16.2%)
	Reverse	0.278	0.204 (73.4%)	0.074 (26.6%)	0.061 (82.5%)	0.013 (17.5%)

Table 11: Decomposition of total input influence for different models in various datasets for context-question (forward) vs question-context (reverse) using Llama-2.

dataset	model	accuracy		influence		
		original	para	context	semantic	linguistic
IMDb	RoBERTa	94.8	94.0	0.472	0.444 (94.2%)	0.028 (5.8%)
	BERT	93.3	92.9	0.483	0.458 (94.7%)	0.025 (5.3%)
Yelp	RoBERTa	94.3	93.9	0.472	0.445 (94.2%)	0.027 (5.8%)
	BERT	92.9	92.6	0.518	0.488 (94.2%)	0.030 (5.8%)
Amazon	RoBERTa	91.0	89.5	0.361	0.325 (90.0%)	0.036 (10.0%)
	BERT	91.2	90.3	0.425	0.389 (91.5%)	0.036 (8.5%)
SST-2	RoBERTa	87.4	82.5	0.210	0.171 (81.4%)	0.039 (18.6%)
	BERT	89.0	84.7	0.274	0.229 (83.5%)	0.045 (16.5%)
TweetEval	RoBERTa	85.2	74.5	0.570	0.469 (82.2%)	0.101 (17.8%)
	BERT	77.7	75.3	0.592	0.506 (85.5%)	0.086 (14.5%)

Table 12: Decomposition of total input influence for different models in various sentiment classification dataset

accuracy		influence		
original	para	response	response-semantic	response-linguistic
79.2	64.4	0.399	0.283 (70.9%)	0.116 (29.1%)

Table 13: Influence from semantic meaning and linguistic realization of the responses in grade classification task in Hewlett dataset.