Hype or not? Formalizing Automatic Promotional Language Detection in Biomedical Research

Bojan Batalo[1], Erica K. Shimomoto[1], Dipesh Satav[2], Neil Millar[2],
[1]National Institute of Advanced Industrial Science and Technology (AIST),
[2]University of Tsukuba,
{bojan.batalo,kidoshimomoto.e}@aist.go.jp,
dsatav@cvlab.cs.tsukuba.ac.jp, millar.neil.gm@u.tsukuba.ac.jp

Abstract

In science, promotional language ('hype') is increasing and can undermine objective evaluation of evidence, impede research development, and erode trust in science. In this paper, we introduce the task of automatic detection of hype, which we define as hyperbolic or subjective language that authors use to glamorize, promote, embellish, or exaggerate aspects of their research. We propose formalized guidelines for identifying hype language and apply them to annotate a portion of the National Institutes of Health (NIH) grant application corpus. We then evaluate traditional text classifiers and language models on this task, comparing their performance with a human baseline. Our experiments show that formalizing annotation guidelines can help humans reliably annotate candidate hype adjectives and that using our annotated dataset to train machine learning models yields promising results. Our findings highlight the linguistic complexity of the task and the potential need for domain knowledge. While some linguistic works address hype detection, to the best of our knowledge, we are the first to approach it as a natural language processing task. Our annotation guidelines and dataset are available at https://github.com/hype-busters/eacl2026-hype-dataset.

1 Introduction

Detecting AI-generated content, plagiarism, fake news, and bias related to health topics has become the target of many machine learning systems to ensure evidence-grounded information is correctly communicated to the general audience. However, one issue yet to be addressed is the increasing use of promotional language – a phenomenon referred to as 'hype' (Millar et al., 2019).

For example, investigators promote the significance and novelty of their research using exaggerated terms (*revolutionary*), dramatically describe research problems (*daunting*), amplify the scale and rigor of their methods (*extensive*, *robust*), and the utility of the results (*actionable*, *impactful*). Increasing use of hype has been demonstrated in biomedical funding applications (Millar et al., 2022a) and journal publications (Vinkers et al., 2015), while comparable trends are evident in other research fields (Weidmann et al., 2018).

A **thorough** investigation was conducted on the papers at hand. **Meticulous** selection with **stringent** qualifying criteria resulted in ten **excellent** publications.

Our study **clearly** shows that *meticulous* hemostasis before closure is **essential** for management of patients.

Figure 1: Examples of 'hype' such as gratuitous amplification of research rigour; determining hype depends on the context, as exemplified by adjective 'meticulous'.

Hype in science is a cause for concern. As the former editor-in-chief of JAMA Network journals points out, words such as *groundbreaking*, *transformative*, or *unprecedented* are rarely justified and may undermine objective assessment, impeding the development of further studies, policies, clinical practice, and knowledge translation (Bauchner, 2023). Moreover, promotional and confident language can bias readers' evaluation of research (Van den Besselaar and Mom, 2022; Peng et al., 2024), and public trust in science is eroded when promotional language creates unrealistic expectations or misrepresents findings (Intemann, 2022).

Previous linguistic studies have sought to detect hype language through collecting and analyzing raw corpora of biomedical research texts (Millar et al., 2019, 2022a, 2023), and have identified a lexicon of 140 adjectives that can carry promotional meaning in biomedical texts (Millar et al., 2022a,b). Although this lexicon can help identify candidates, discussions about whether a specific term constitutes hype remain problematic, as a given word or phrase may be promotional depending on the context. For instance, adjectives like *essential* and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6979–6992

meticulous can promote significance or rigor, but they may also occur in a neutral context or technical phrase (e.g., essential fatty acid, meticulous hemostasis), as shown in Figure 1.

Therefore, we introduce the task of automatic detection of hype language, focusing on biomedical research texts. Specifically, we are concerned with the detection of hype in research writing from a linguistic point of view, and not fact-checking the claims made by the authors. To the best of our knowledge, we are the first to address it as a natural language processing (NLP) task. As a starting point, and highlighting that this is not an exhaustive definition, we follow Millar et al. (2019) and define the concept of hype as 'hyperbolic and/or subjective language that authors use to glamorize, promote, embellish, and/or exaggerate aspects of their research.' Likewise, we consider their lexicon of adjectives as a starting point, as adjectives are prototypical and the most common means of expressing evaluation (Martin and White, 2003).

We first address data annotation. Determining whether a word is used with promotional intent involves subjective judgment based on context and interpretation. Moreover, the general definition of hype may be inadequate for distinguishing ambiguous cases, forcing annotators to rely on intuition. Therefore, we propose formal annotation guidelines to determine whether an adjective is used in a promotional manner based on its semantics, function, and context. Using these guidelines, we then manually annotate a dataset of 1270 sentences from the NIH grant application abstract corpus containing potentially promotional adjectives.

Finally, we formulate automatic hype detection as a text classification task, and test traditional NLP classifiers using bag-of-words and word embeddings as features, as well as finetune small-scale language models, such as BERT and GPT-2, and use off-the-shelf large language models such as LLAMA3.1-INST. and GPT-4O-MINI. We compare their performance to a human baseline that did not know our guidelines. Our results indicate that formalizing annotation guidelines helps humans reliably annotate Hype adjectives, and training NLP models on the annotated dataset yields promising results for automatic detection of hype.

2 Related Works

To the best of our knowledge, no prior work within NLP has been done on promotional language use in scientific publishing. However, here we highlight some work relevant to the analysis of such language in various publication outlets.

Tackling promotional language Bhosale et al. (2013) propose a model for detecting promotional content in Wikipedia articles. They use both n-gram language models and probabilistic context-free grammars and show that incorporating stylistic and syntactic features outperforms models trained only on lexical and meta-features.

In addition, several works examine exaggerated claims and statements related to health news. Li et al. (2017) analyze the discrepancies in the strength of claims made in scientific journal articles, and news articles in which they are reported to the public, intending to detect media manipulation; the authors use a bag-of-words approach with lexical features to train simple classifiers. Patro and Baruah (2021) expand on the same task and dataset, and utilize BERT (Devlin et al., 2019) to extract relation phrases, used to train classifiers to recognize the strength of a claim made in the journal article and its corresponding news reporting, which are compared. A similar task is undertaken by Wright and Augenstein (2021), who analyze pairs of journal articles and their press releases, as well as by Kamali et al. (2024), who focus on persuasive writing for health disinformation.

Works on exaggeration focus on the amplification of relational claims between a source and the reporting text (X *might cause* cancer vs. X *causes* cancer). These studies primarily examine modifications of certainty and confidence (commonly referred to as epistemic stance) typically expressed through modal verbs (*might*, *will*, *can*), hedges (*likely*, *clearly*), and causal triggers (*causes*, *leads to*). Our work on hype is primarily concerned with value judgments (*innovative*, *promising*, *impactful*). While the intensity of hype can be modified by epistemic stance features, these are not central to its identification.

Understanding underlying intentions Simultaneously, there is a growing interest in understanding the underlying intentions behind utterances, exemplified by tasks such as sentiment analysis (Zhang et al., 2023), emotion analysis (Del Arco et al., 2024), and stance detection (Mu et al., 2024). Sentiment analysis, one of the most prominent NLP tasks, identifies whether a writer's orientation towards an object, person, or idea is positive, neutral, or negative. Closely related to this is emotion anal-

ysis, which focuses on specific emotions, whereas the study of stance examines attitudes, judgments, and commitment.

We find similarities between hype detection and sentiment analysis, especially in its fine-grained version, as words with similar meanings can vary in promotional intensity (e.g., *new* vs. *novel* vs. *innovative* vs. *groundbreaking*). However, while sentiment analysis deals with only two polarities of sentiment, i.e., positive and negative, hype language can be used to promote different aspects of research (Millar et al., 2022a), and, therefore, it relates more closely to emotion detection. Also, just like emotion analysis and stance detection, hype detection is highly context-dependent; a given word or phrase may be promotional in one setting while not in others.

Furthermore, hype detection is tied to identifying authors' intention to promote aspects of their research. Such intention can be expressed through different choices of words, connecting our task to lexical choice (Ding et al., 2021), which aims at choosing words to communicate the intended information to the reader in a generated text.

Nevertheless, a key distinction of our task compared to the aforementioned approaches is that it requires domain knowledge of the specific research field — in this case, biomedical research, as some words or phrases may be part of a technical term rather than used to make value judgments.

3 NIH grant application abstract corpus

The starting point of our work is the corpus of raw texts compiled by Millar et al. (2022a), which comprises 901,717 abstracts of successfully funded grant applications submitted to and approved by the National Institutes of Health (NIH) in the United States of America. The NIH corpus has 335 million words, out of which 36.4 million are adjectives. Millar et al. (2022a) give the following broad definition of hype: *"If the adjective has **positive value judgment**, and can be **removed or replaced without loss in meaning**, it is potentially hype"*. Following this definition, they have identified 140 adjectives as 'potentially hype' in the NIH corpus.

These adjectives were categorized in eight groups, based on the aspects of research they are promoting: **importance** (*critical*, *fundamental*), **novelty** (*innovative*, *groundbreaking*), **rigor** (*rigorous*, *robust*), **scale** (*comprehensive*, *interdisciplinary*), **utility** (*generalizable*, *transformative*), **quality** (*exceptional*, *prestigious*), **attitude** (*interesting*, *remarkable*), **problem** (*daunting*, *unmet*).

As a first step, we select the **novelty** and **rigour** groups of adjectives, which emphasize the novelty and innovation of proposed research as well as its rigour and quality, aspects of research reinforced by the academic peer-review system and the competitive funding process. The **novelty** group comprises 11 members: *creative*, *emerging*, *first*, *groundbreaking*, *innovative*, *latest*, *novel*, *revolutionary*, *unique*, *unparalleled*, *unprecedented*; while the **rigour** group comprises 15 members: *accurate*, *advanced*, *careful*, *cohesive*, *detailed*, *nuanced*, *powerful*, *quality*, *reproducible*, *rigorous*, *robust*, *scientific*, *sophisticated*, *strong*, *systematic*.

To compile our dataset, we use CQPweb (Hardie, 2012) to search for novelty adjectives through the recent NIH corpus abstracts, covering years from 2016 to 2020. This search yielded a total of 161,469 occurrences, covering 84,299 abstracts. Due to time and resource constraints, we limit ourselves to a smaller corpus, which can be annotated and manually examined. We randomly choose 50 sentences per adjective, resulting in 1,300 samples to be annotated, drawn from 1,287 abstracts.

4 Proposed annotation guidelines

We designed the initial annotation guidelines with the help of a linguist expert in promotional language and corpus linguistics. The guidelines comprise several sequential steps, which might require high proficiency in English. They are designed to be easy to follow, and each step provides positive and negative examples.

Determining whether an adjective carries promotional intent requires considering the context in which it is used. Therefore, we chose to evaluate adjectives within their full sentential context, while context outside of the sentence is not considered. We make this choice as sentences are the most common unit of expressing a claim. While checking if a claim is true or not, i.e., fact-checking, may require access to several sentences in the text (e.g., access to the results portion that justifies the claim) or even external sources, we observed that one can generally infer if an adjective carries promotional meaning within the sentence context. With this starting point, the steps of the annotation guidelines are defined as follows.

Step 1: Value-judgment. Determine whether the adjective implies positive value judgment. Most

of the selected adjectives do; this includes priority claims ("*first* method to..."). If yes, proceed to steps 2-6. If not, the adjective is deemed as `Not Hype`. Such cases are typically acronyms, technical/domain-specific meaning, literal meaning or part of a proper noun phrase ("*first* step", "*Creative* Scientist, Inc. (CSI)".

Step 2: Hyperbolic. If the adjective is hyperbolic or exaggerated, it is deemed as `Hype`. This is a relatively unambiguous category which contains, but is not limited to, a pre-determined set of adjectives: *revolutionary*, *unprecedented*, *unparalleled*, *groundbreaking*. In the great majority of cases, these adjectives are used in a promotional manner ("*unparalleled* opportunity", "*revolutionary* tool").

Step 3: Gratuitous. Determine whether the adjective adds little to the proposition. If removed, and the propositional content and structural integrity of the sentence remain unchanged (typically when adjective is used in attributive relationship, e.g., "This *cohesive* platform will encourage scientists..."), or the adjective represents a tautology and is redundant ("we *discovered* a *novel* gene"), the adjective is deemed as `Hype`. If, however, the propositional content would be substantially altered ("this treatment uses a *novel* approach"), or justification is given within the sentence (typically when adjective is used in predicative relationship, e.g., "the study is *innovative* because no previous research..."), the adjective is `Not Hype`.

Step 4: Amplified. If the strength of the adjective is amplified through the use of modifiers, the adjective is deemed as `Hype` ("truly *novel*", "extremely *rigorous*", "highly *cohesive*").

Step 5: Coordinated. If the adjective is coordinated with one or more hype candidate adjectives, the adjective is deemed as `Hype` (e.g., "*innovative* and *strong* researcher"). We also term this phenomenon *adjective stacking*.

Step 6: Broader context. When ambiguous, consider whether the sentence in question contains other instances of potential hype. If the overall tone of the sentence is promotional (such as when multiple potential hype adjectives are used), the adjective is deemed as `Hype` (e.g., "The faculty has an outstanding track record of *creative* and innovative research, advanced mentoring, and robust research funding, and thus attracts outstanding applicants."). This step is the most subjective and potentially requires discussion and agreement of several annotators to maintain consistency of annotation.

The proposed guidelines offer some formal definitions of what constitutes linguistic hype, and can help a human annotator determine whether an adjective is hype, depending on the context. In some cases, the guidelines may prove insufficient and require further discussion. Detailed examples of the annotation guidelines are given in §A.

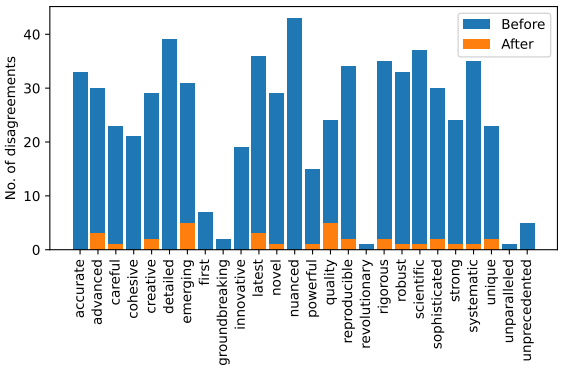

Figure 2: Disagreements between annotators, before and after the discussion. Disagreements were largely resolved; some remained mainly on *emerging*, *latest*, *quality* and *advanced*.

4.1 Annotation process

Four voluntary researchers annotated the dataset, in different stages of academic career (graduate student, post-doctoral researcher, researcher, university Professor), where only one is a native English speaker. They were verbally instructed to follow the proposed guidelines, knowing that the annotations would be used for scientific publication. This process followed the proposed guidelines step-by-step. The annotators are shown a sentence containing a highlighted adjective. Following the guidelines, the annotators first look at the adjective in question and then consider the context of the entire sentence.

In principle, an adjective can be deemed `Hype` while satisfying multiple criteria, e.g., a sentence can satisfy **hyperbolic**, **gratuitous** and **coordinated** rationales at the same time; therefore, the annotators are encouraged to include every criterion that has been met.

While none of the annotators are experts in biomedical data, we have consulted with a biomedical expert about the potential hype adjectives considered in this study, who has provided several examples of the use of such adjectives in biomedical terms that do not constitute hype. For example:

- "This *unique* identifier will also allow tracking ..." (`Not Hype`)

- "These human lab models provide a *unique* opportunity to provide a deeper understanding of the alcohol - response phenotypes." (Hype, gratuitous)

- "Vaccination with any of the three antigens elicits a *robust* immune response that protects from disseminated." (Not Hype)

- "Taken together, our proposed use of these three assays in this supplement provide a *robust* approach to test the pitfalls outlined above." (Hype, gratuitous)

Such examples were incorporated in the annotation guidelines provided to the annotators.

The initial stage resulted in fair amounts of disagreement, as indicated by the pairwise Cohen's Kappa values ranging from 0.30 to 0.52, shown in Table 1. A discussion session was held to resolve conflicts and reevaluate annotation guidelines. The disagreement level differs for each adjective; as can be seen in Figure 2.

Disagreements arose from either misunderstanding guidelines, accepting common academic parlance as not promotional, or confusing the tasks of detecting linguistic promotion with fact checking, e.g. "*unparalleled* results", should be labeled as Hype even if true. This was clarified during discussions, which resolved most disagreements, by strictly adhering to the guidelines and the provided examples in particular.

For adjectives corresponding to the *hyperbolic* guideline, disagreements were minimal, as almost all applications of the guideline resulted in a Hype decision. Similarly, adjectives such as *powerful* and *innovative* are often unambiguously Hype. *First* is often clearly either a priority claim or used as a numbering device, e.g., "*first* weeks of therapy."

Adjectives from the **rigour** group, such as *nuanced* and *detailed*, caused the most disagreement, e.g., "*nuanced* approach", "*detailed* analysis", mostly because they are common in academic writing; however, they often fall under the *gratuitous* criterion and can be removed without any loss in meaning. The same holds for *scientific* and *sophisticated*, which are most often promoting rigour without adding substance.

From the **novelty** group, adjectives such as *emerging* and *latest* were the most difficult, and the guidelines proved insufficient to fully categorize them. They are often used to refer to *emerging* phenomena when establishing context for the proposed research (e.g., "In the past decade, *emerging* neuroimaging technique" vs. "represent a significant *emerging* worldwide health threat") or the *latest* publications presented at a scientific conference (e.g., "This meeting is focused on the *latest* findings..." vs. "...session on applications of the *latest* new technologies"); in these cases, it is necessary to look at the broader context within the sentence to determine if their use is promotional. Examples of difficult cases are shown in Table 2.

	A	**B**	**C**	**D**
A	–	0.52 (0.96)	0.52 (0.94)	0.30 (0.94)
B	0.52 (0.96)	–	0.49 (0.96)	0.37 (0.95)
C	0.52 (0.96)	0.49 (0.95)	–	0.42 (0.96)
D	0.30 (0.94)	0.37 (0.95)	0.42 (0.96)	–

Table 1: Pairwise Cohen's Kappa between the annotators A, B, C and D. Adjusted agreement values after the discussion stage are displayed in brackets.

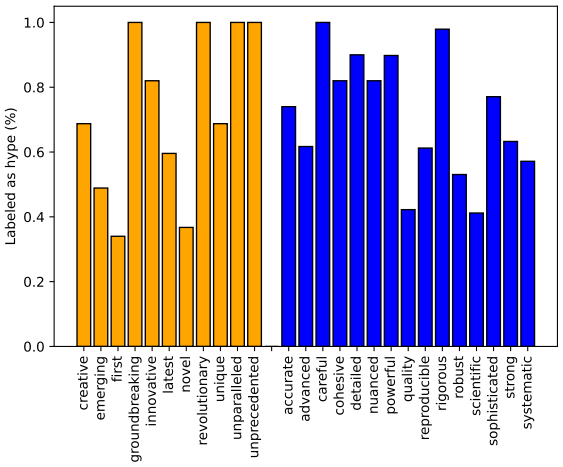

Figure 3: Percentage of samples labeled as 'hype' in the final dataset.

Discussion resolved 545 out of the 575 disagreements, yielding the final dataset with 1270 annotated sentences. 30 samples were set aside; they highlight the difficulties of designing broad guidelines, especially for adjectives *emerging*, *advanced*, *quality*, *latest*, *unique*, and *creative*, and will be further analyzed to improve guidelines in the future.

4.2 NIH hype dataset

The annotation resulted in a dataset of 1270 unique sentences with potential hype adjectives, 917 of which are deemed as Hype, and 353 Not Hype. The percentage per adjective is shown in Figure 3.

Five adjectives are Hype in 100% of the cases –

Category	Adjective	Example Sentence	Assessment
Novelty	First	"In the project's **first** aim we address the role of B cell activation and autoantibody formation in pain, functional and neuropsychiatric outcomes after fracture injury in mice."	`Not Hype`
		"Our proposed studies are some of the **first** to champion the notion that the 'triangulation' of disparate scientific studies and discoveries [...]."	`Hype`
	Latest	"We will determine whether there is a minimum fraction of cyst cells that need to be targeted by reactivation to reverse ADPKD and determine the **latest** disease stage at which ADPKD retains reversibility."	`Not Hype`
		"SCMM prides itself on bringing a critical mass of medical mycology researchers together from the south central region of the United States to share the **latest** research being conducted in the various labs."	`Hype`
Rigour	Scientific	"Consequently, the aims of the Administrative Core are to (1) provide **scientific** and administrative leadership."	`Not Hype`
		"The identification of such brain-specific markers will likely open new avenues for **scientific** investigation of BD."	`Hype`
	Nuanced	"The rationale for this project is that an understanding of **nuanced** sleep-based consolidation and the role of emotion beyond the restricted cohort of young adults will foster the ability to determine the underlying age-related changes in the phenomenon."	`Not Hype`
		"More **nuanced** understanding of how individual parent own obesogenic eating behaviors and concern about child weight dyadically interact to shape their own and their partners' child feeding practices may be critical in order to design and implement intervention strategies [...]."	`Hype`

Table 2: Examples of cases difficult to annotate.

Rationale	Novelty	Rigour	Total
Hyperbolic	205	3	208
Gratuitous	180	503	683
Amplified	7	15	22
Coordinated	49	89	138
Broader context	145	75	220

Table 3: Distribution of rationales for 'hype' classification following the annotation guidelines.

groundbreaking, *revolutionary*, *unparalleled*, *unprecedented*, and *careful*. Most of these adjectives correspond to the **hyperbolic** category and are extreme examples of hype. Similarly, *rigorous*, *powerful*, and *detailed* are almost always deemed as **gratuitous** hype. Adjectives such as *innovative*, *creative*, and *unique* are more often than not used in a promotional manner, with 82%, 68%, and 68% of their uses constituting hype, respectively. Remaining adjectives are more varied in their use and denote hype when used in a promotional context.

This distribution is reflected in annotator rationales, shown in Table 3 (for the full breakdown by adjective, we refer the reader to §D). The most numerous category is **gratuitous** (683 samples) with adjectives such as *rigorous*, *careful*, and *nuanced* denoting scientific rigour and *innovative*, *creative*, *unique* denoting its novelty. **Broader context** (220 samples) is a subjective category based on how promotional the overall sentence is, as opposed to the literal meaning and the near context of the adjective itself. Adjectives corresponding to the **hyperbolic** category account for almost half of the hype examples of novelty (205 samples).

The least prominent is the **Amplified** category (22 samples). This may reflect writers' tendency to avoid overtly promotional language ("highly innovative"), since amplifying adjectives adds a layer of subjective evaluation. Similarly, the **coordinated** category comprises only 138 samples.

5 Preliminary experiments

We frame hype detection as a text classification task: given a set of training sentences $S = \{s_i\}$, with known classes $C = \{\texttt{Hype}, \texttt{Not Hype}\}$, classify an input sentence s_q as one of the classes in C. We evaluate traditional text classification methods, pre-trained language models (PLMs), and large language models (LLMs) to assess task complexity. Finally, we obtain a preliminary human baseline to judge the construct validity of our annotation

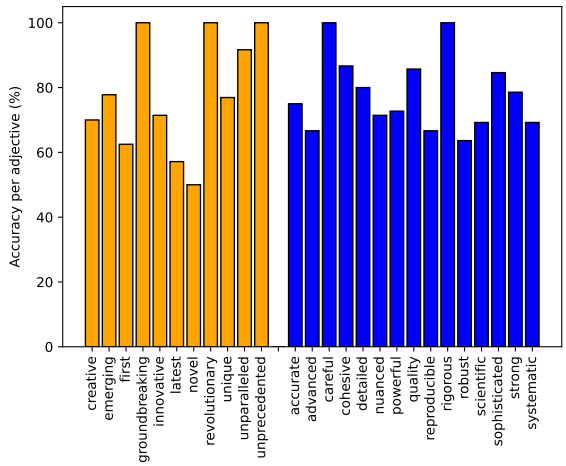

Figure 4: Accuracy per adjective for SVM.

guidelines.

The dataset was split into development and hold-out test sets in an 8:2 ratio, stratified by class. The exact percentage of samples per adjective in the training set is shown in §D. Performances for all models are reported on the hold-out test set.

For traditional classifiers and PLMs, hyper-parameter search was done on the development set with 10-fold cross-validation. LLMs were prompted whether an adjective is used in a promotional manner, given the sentence and the broad definition of hype (Millar et al., 2022a) (+Broad Prompt), and a more strict definition (+Strict Prompt). A verbalizer mapped the spans of "HYPE" and "NOT HYPE" in the response to appropriate labels. Full prompts are given in §B.

For the human baseline, we recruited two researchers with 10+ years of academic experience to manually classify the hold-out test set, provided only with the broad definition of hype. Neither LLMs nor human evaluators have access to the annotation guidelines - only the sentence to be evaluated and the broad definition of hype.

We evaluate using standard classification metrics, i.e., accuracy, precision, recall, and F1-score. Additionally, we report macro F1-score to account for class imbalance. For the human baseline, we report average scores of the two subjects.

5.1 Traditional Text Classification Methods

We train the following traditional text classification methods: Multinomial Naive Bayes (MNB), Multivariate Bernoulli Naive Bayes (MVB), Latent Semantic Analysis (LSA), and Support Vector Machines with a linear kernel (SVM). We consider bag-of-words of unigrams and the averaged word embedding obtained via pre-trained GLOVE[1]. We include MAJORITY, a classifier that always predicts Hype, to establish a lower-bound baseline.

Method	Feature	A	P_+	R_+	$F1_+$	M-F1
MAJORITY	-	0.722	0.722	1.000	0.838	0.419
HUMAN	-	0.767	**0.886**	0.783	0.824	**0.730**
MNB	BOW	0.773	0.781	0.951	**0.858**	0.645
MVB		0.741	0.748	**0.967**	0.844	0.547
LSA		0.639	0.756	0.739	0.747	0.559
SVM		**0.776**	0.795	0.929	0.857	0.672
SVM	GLOVE	0.761	0.783	0.924	0.848	0.644

Table 4: Performance of traditional text classifiers in terms of Accuracy (A), Precision (P_+), Recall (R_+), F1-score ($F1_+$), and Macro F1-score (M-F1). Precision, Recall and F1-score considers Hype as the class of interest. Best values are highlighted in bold.

Results are in Table 4. All methods had similar performance in terms of accuracy, with MNB and SVM capturing more Not Hype samples and having better recall and macro F1-scores. While LSA achieved lower accuracy compared to MAJORITY, its higher macro F1-score indicates it is identifying the Not Hype cases to some extent. SVM achieved the best results in terms of accuracy, but is still behind the human baseline overall, with a lower macro F1-score. This is likely because adjectives such as *groundbreaking*, *revolutionary*, and *careful* mostly appear labeled as Hype, and possibly bias the results towards only classifying sentences as Hype, as observed in Figure 4.

Interestingly, methods based on BOW were capable of performing reasonably well compared to GLOVE. This may be due to using average GLOVE word embeddings with the SVM classifier and potentially losing context. The overall results indicate hype detection may require more complex language modeling than BOW features and word embeddings, as it is highly context-dependent.

5.2 Language Models

Given the size of our dataset, we chose to finetune smaller PLMs, such as BERT[2] and DISTILBERT[3]. We also finetuned a model pre-trained on biomedical data[4], BIOMEDBERT, to see if prior knowledge of the biomedical field would affect performance.

[1] glove.42B.300d
[2] https://huggingface.co/google-bert/bert-base-uncased
[3] https://huggingface.co/distilbert/distilbert-base-uncased
[4] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

Finally, we finetuned a comparable-size causal language model, GPT-2[5]. For all models, we added a classification head and performed two training schemes: 1) freezing the PLM and training only the classification head (❄), and 2) finetuning the PLM along with the classification head (+Finetuning).

We used AdamW optimizer with HuggingFace's default hyperparameters, varying only the weight decay and learning rate through a 10-fold cross-validation on the development set, with learning rate values of $[1e-5, 2e-5, 3e-5]$, weight decay values of $[0.0, 0.01, 0.1]$, and batch size of $[8, 16, 32]$. We set the number of epochs to 30, but also implemented early stopping with a patience of 3. Most models did not train past 10 epochs. The best hyperparameters found are reported in §C. Finally, we also experimented with two LLMs, LLAMA3.1-INST.[6] and GPT-4O-MINI.

Table 5 shows the results. We report performance on the whole dataset and a breakdown for each adjective group. We can see that all PLMs performed poorly when frozen, falling behind the traditional methods. Their high precision, recall, and F1-scores for Hype but low macro F1-scores show that while they correctly detect Hype instances, they fail to detect Not Hype instances.

Finetuning leads to substantial performance improvements, surpassing our HUMAN baseline, with BERT-based models performing the best. We believe this is likely because they are bidirectional and, therefore, have a better overall understanding of the entire sentence. This result highlights the importance of the surrounding context of the potential hype words, as the rationale might be expressed before or after the word.

Moreover, we can see that all models performed worse on **rigour** than on **novelty** data. Among finetuned models, GPT-2 had the largest gap between its performances on each adjective group, struggling to detect Not Hype instances for Rigour.

BIOMEDBERT prior knowledge on medical data does not seem to be of help, as DISTILBERT showed a more balanced performance among both categories without domain-specific knowledge, leading to the best Macro F1 score overall.

Zero-shot classification with LLMs led to the lowest performances, with LLAMA3.1-INST. performing worse than frozen PLMs. Even the best performing LLM, GPT-4O-MINI, fell behind the HUMAN baseline. While they achieved decent precision rates for both types of prompts, the low recall and low F1-scores indicate that they are biased towards classifying inputs as Not Hype, i.e., a high number of false negatives.

This is likely due to the broad definition of hype (+Broad Prompt), under which the model can rely on inherent biases from the training data. Many promotional adjectives have become very common in the scientific parlance and, therefore, might not seem out of place. When asked to adhere more strictly to the definition of hype (+Strict Prompt), the models changed their decision from Not Hype to Hype for some sentences, leading to decreased precision, but increased recall and macro F1-scores.

Finally, we can see that LLMs presented the highest gap between their performances on each adjective group, with LLAMA3.1-INST. failing to detect both instances of Hype and Not Hype for **rigour**.

5.3 HUMAN Baseline

Table 4 shows that a blind HUMAN baseline aligns reasonably well with the proposed guidelines. In terms of precision and macro-F1 score, HUMAN baseline performs better than traditional text classifiers, indicating that HUMAN Hype predictions are often correct; however, similarly to LLMs, human subjects have a stronger tendency to predict Not Hype, resulting in lower recall. This bias is likely formed due to their exposure to scientific writing and accepting academic parlance, even if promotional, as commonplace.

Such bias had to be resolved during the construction of annotation guidelines as well, and can also be seen in LLMs. However, when compared to LLMs prompted with the same instructions, humans are much better at detecting promotional language, and especially when it comes to adjectives denoting **rigour**, as can be seen in Table 5.

The HUMAN baseline showed that hyperbolic words such as *revolutionary* might not always be perceived as Hype. One might argue that stating a method is *revolutionary* does not constitute hype if this method was, in fact, *revolutionary* to the research field. Alternatively, even if the method truly is revolutionary, calling it so may still be considered promotional. In a scientific context, such evaluative language can impose a judgment on the reader rather than letting the research speak for itself, thus raising questions about objectivity and,

[5]https://huggingface.co/openai-community/gpt2

[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Method	Overall					Novelty					Rigour				
	A	**P_+**	**R_+**	**$F1_+$**	**M-F1**	**A**	**P_+**	**R_+**	**$F1_+$**	**M-F1**	**A**	**P_+**	**R_+**	**$F1_+$**	**M-F1**
HUMAN	0.767	0.886	0.783	0.824	0.730	0.745	0.902	0.741	0.800	0.712	0.782	0.876	0.814	0.840	0.744
BERT❄	0.721	0.732	0.967	0.833	0.489	0.731	0.740	0.974	0.841	0.481	0.714	0.726	0.961	0.827	0.493
+Finetuning	**0.862**	0.866	0.956	**0.909**	0.812	0.879	0.892	0.949	0.920	0.837	**0.850**	0.848	0.961	**0.901**	0.793
DISTILBERT❄	0.752	0.759	0.961	0.848	0.585	0.768	0.775	0.962	0.858	0.608	0.741	0.748	0.961	0.841	0.568
+Finetuning	0.858	**0.893**	0.913	0.903	**0.821**	0.879	**0.912**	0.924	0.918	0.845	0.843	**0.879**	0.904	0.892	**0.804**
BIOMEDBERT❄	0.713	0.719	**0.989**	0.832	0.416	0.722	0.728	**0.987**	0.838	0.419	0.707	0.712	**0.990**	0.828	0.414
+Finetuning	0.835	0.858	0.923	0.890	0.780	**0.888**	0.903	0.949	**0.925**	**0.851**	0.795	0.826	0.904	0.863	0.729
GPT-2❄	0.717	0.725	0.978	0.833	0.455	0.740	0.742	**0.987**	0.847	0.486	0.700	0.713	0.971	0.822	0.433
+Finetuning	0.807	0.832	0.918	0.873	0.737	0.879	0.892	0.949	0.920	0.837	0.755	0.789	0.895	0.839	0.662
LLAMA3.1-INST.															
+Broad Prompt	0.451	0.879	0.277	0.421	0.450	0.639	0.885	0.582	0.702	0.622	0.313	0.833	0.048	0.090	0.269
+Strict Prompt	0.482	0.750	0.424	0.542	0.474	0.657	0.776	0.747	0.761	0.577	0.354	0.679	0.181	0.286	0.348
GPT-4O-MINI															
+Broad Prompt	0.631	0.852	0.592	0.699	0.612	0.833	0.851	0.937	0.892	0.766	0.483	0.854	0.333	0.479	0.483
+Strict Prompt	0.675	0.789	0.750	0.769	0.610	0.796	0.788	**0.987**	0.876	0.649	0.585	0.789	0.571	0.663	0.562

Table 5: Performance of language models in terms of Accuracy (A), Precision (P_+), Recall (R_+), F1-score ($F1_+$), and Macro F1-score (M-F1). Precision, Recall and F1-score considers Hype as the class of interest. Best values are highlighted in bold.

indeed, shifting values. Therefore, we separate our proposed task from fact-checking and focus on detecting promotional language from a purely linguistic point of view.

Looking closer at the misclassified sentences, we see that the HUMAN baseline struggles with adjectives such as *first*, *latest*, *emerging*, *scientific*, and *nuanced*. Many of these adjectives were among the most difficult words to annotate in our dataset, as mentioned in Section 4.1. Additionally, human subjects often do not classify *first* as hype, while we do define it as such due to priority claims. These results, combined with the human bias towards common academic parlance, highlight the need for strict annotation guidelines.

6 Conclusion

To the best of our knowledge, this is the first attempt at identifying promotional language in scientific texts. We developed formal annotation guidelines and applied them to a set of texts from the NIH archive. Multiple machine learning models were used for the defined text classification task, determining whether a sentence containing a potentially promotional adjective is Hype or Not Hype. Our results suggest that formalizing annotation guidelines helps humans reliably annotate potentially hype adjectives, and that using such an annotated dataset to train machine learning models yields promising results. We highlight the complexity of the task and the potential impact of domain knowledge, not only for machine learning algorithms, but for humans as well.

Progress in this task requires refinement of our annotation guidelines, engineering, and expansion to other types of academic texts. For future work, we plan on developing guidelines for other groups of hype adjectives besides **novelty** and **rigour**, and greatly expanding the lexicon, possibly using automatic expansion techniques, to include adverbs, nouns, and verbs, among others. We also plan on scaling up both the data annotation and human evaluation processes to better understand the perception of promotional language in the scientific community. Additionally, we would like to consider the intensity of individual adjectives; not every adjective carries the same intensity of hype, and taking this into account would allow us to expand the hype detection beyond the binary classification task.

Finally, we are particularly interested in applying the current work to downstream tasks, such as automated tools for flagging and editing of promotional language in scientific writing, and as a potentially starting point for scientific fact-checking.

Limitations

This work proposed the task of automatically detecting hype language in biomedical research. While our proposed annotation guidelines and preliminary experiments on annotated data follow-

ing these guidelines have shown promising results, there are important limitations worth noting:

- The perception of hype is subjective. While we tried to approach this task logically, we cannot ignore our inherent biases when creating the guidelines and annotating the data.
- We only produced annotation guidelines for two of the eight adjective groups within the NIH grant application abstract corpus. We expect to face similar challenges when developing guidelines for the other groups, but we cannot guarantee that we will achieve similar results.
- Our proposed dataset is limited to 1270 sentences from a corpus of almost a million abstracts. Experiments and analysis on such a small dataset might not be representative of a general distribution.
- We focus only on biomedical research. While academic writing has a distinct style, hype language in other fields is likely expressed using different words and at different intensity levels.
- We only tackled the detection of hype for English. Studying promotional language in scientific publishing in other languages might require different approaches.
- Experiments shown in this work were performed using a 16-GB NVIDIA V100 GPU. We reckon that experimenting with more recent LLMs will require larger computing capacity.

Ethics Statement

We collected and annotated data for this study in accordance with ethical research standards. All participants provided informed consent prior to participation. Annotators for our dataset were volunteers who generously dedicated their free time to assist us. For the human baseline evaluation, the two recruited researchers were compensated for their time at rates above local minimum wage standards. The data contain no personally identifiable information, and all experiments comply with the terms of service of the data sources.

Our goal is to mitigate the use of hype language in scientific publishing in biomedical research. However, one may misuse the findings of this paper to purposely include hype language.

Acknowledgements

This paper was supported by grant No. 25K00851 from the Japan Society for the Promotion of Science. Additionally, this paper is based on results obtained from project JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Howard Bauchner. 2023. Hype, the responsibility of authors and editors, and the subjective interpretation of evidence. *JAMA Network Open*, 6(12):e2349125–e2349125.

Shruti Bhosale, Heath Vinicombe, and Raymond Mooney. 2013. Detecting promotional content in Wikipedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1851–1857, Seattle, Washington, USA. Association for Computational Linguistics.

Flor Miriam Plaza Del Arco, Alba A Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710.

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

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations*.

Andrew Hardie. 2012. Cqpweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3):380–409.

Kristen Intemann. 2022. Understanding the problem of "hype": Exaggeration, values, and trust in science. *Canadian Journal of Philosophy*, 52(3):279–294.

Danial Kamali, Joseph D. Romain, Huiyi Liu, Wei Peng, Jingbo Meng, and Parisa Kordjamshidi. 2024. Using persuasive writing strategies to explain and detect health misinformation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17285–17309, Torino, Italia. ELRA and ICCL.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Neil Millar, Bojan Batalo, and Brian Budgell. 2022a. Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985-2020. *JAMA network open*, 5(8):e2228676–e2228676.

Neil Millar, Bojan Batalo, and Brian Budgell. 2022b. Trends in the use of promotional language (hype) in national institutes of health funding opportunity announcements, 1992-2020. *JAMA Network Open*, 5(11):e2243221–e2243221.

Neil Millar, Bojan Batalo, and Brian Budgell. 2023. Promotional language (hype) in abstracts of publications of national institutes of health–funded research, 1985-2020. *JAMA Network Open*, 6(12):e2348706–e2348706.

Neil Millar, Francoise Salager-Meyer, and Brian Budgell. 2019. "it is important to reinforce the importance of...":'hype'in reports of randomized controlled trials. *English for Specific Purposes*, 54:139–151.

Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2024. Examining temporalities on stance detection towards covid-19 vaccination. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6732–6738.

Jasabanta Patro and Sabyasachee Baruah. 2021. A simple three-step approach for the automatic detection of exaggerated statements in health science news. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3293–3305, Online. Association for Computational Linguistics.

Hao Peng, Huilian Sophie Qiu, Henrik Barslund Fosse, and Brian Uzzi. 2024. Promotional language and the adoption of innovative ideas in science. *Proceedings of the National Academy of Sciences*, 121(25):e2320066121.

Peter Van den Besselaar and Charlie Mom. 2022. The effect of writing style on success in grant applications. *Journal of Informetrics*, 16(1):101257.

Christiaan H Vinkers, Joeri K Tijdink, and Willem M Otte. 2015. Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: retrospective analysis. *Bmj*, 351.

Nils B Weidmann, Sabine Otto, and Lukas Kawerau. 2018. The use of positive words in political science language. *PS: Political Science & Politics*, 51(3):625–628.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10824–10836, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

A Some examples in the annotation guidelines

Guideline 1: Value-judgement. Does the adjective imply positive value judgment?

- **YES** - Most adjectives will imply a value judgement. This includes priority claims:
  - *Our study will be the **first** to ...*
- **NO** - Typically, acronyms, technical/domain-specific meaning, and literal meaning:
  - *To aid these efforts, **Creative** Scientist, Inc. (CSI)...*
  - *Our curriculum emphasizes the development of critical and **creative** independent thinking...*
  - *In the **first** aim we test the hypothesis...*

Guideline 2: Hyperbolic. Is the adjective hyperbolic or exaggerated?

- *YES* - A relatively unambiguous class that can (likely) be pre-determined:
  - *revolutionary; unprecedented; unparalleled; groundbreaking*

Guideline 3: Gratuitous. Is the adjective gratuitous, adding little to the propositional content?

- *YES (1)* - If removed, the propositional content and structural integrity of the sentence would remain basically unchanged (typically when adjective used in attributive relationship).
  - *To address this, we developed 2 **innovative** technologies.*
  - *Delivering SGR interventions via text messaging is an **innovative** way to increase the reach of this cessation intervention...*
- *YES (2)* - Represents a tautology or is redundant?
  - *discovered a **novel** gene...*
- *NO (1)* - If removed the propositional content of the sentence would be substantially altered.
  - *This is a high risk and high impact project that uses a **novel** approach to aggressively treat local - regional disease.*
- *NO (2)* - The sentence gives justification for the claim (typically when adjective used in predicative relationship).
  - *The proposed study is **innovative** because no previous research has identified how MBC...*

Guideline 4: Amplified. Is the strength of the adjective amplified?

- *YES* - The strength of the adjective made stronger through the use of modifiers:
  - *truly novel; highly innovative; completely unique; etc.*

Guideline 5: Coordinated. Is the adjective CO-ORDINATED with other hype candidates?

- *YES* - Adjective is co-ordinated with one or more hype candidates (adjective stacking):
  - *...**innovative** and **creative** leader...*
  - *...**creative**, **collaborative**, and culturally **diverse** translational scientists...*

Guideline 6: Broader context. When ambiguous, consider whether the sentence contains other instances of potential hype or overt amplification.

- *This **transformative** work will be the **first** study to achieve this level of*
- *The faculty has an **outstanding** track record of **creative** and high - profile research , **superb** mentoring , and **robust** research funding , and thus attracts **outstanding** trainees*

B Large Language Model Prompts

In Figures 5 and 6, we share prompts used to prompt the LLAMA3.1-INST. and GPT-4O-MINI models in Section 5.2, with the broad definition of hype, and its stricter variant.

C Best Hyperparameters

We report the best hyperparameters found for fine-tuning the PLMs in Table 6.

D Dataset details

We provide detailed breakdown of rationales for the Hype classification following the annotation guidelines in Table 7. Additionally, we provide the percentages of samples in the training set for both adjective categories in Tables 8 and 9.

Broad Definition Prompt

You are an expert in linguistics, science communication and biomedicine. Following the DEFINITION of hype, determine whether the ADJECTIVE is used in promotional manner in a given SENTENCE. Output HYPE if yes, NOT HYPE if not. Output only the decision. Do not output reasoning.

DEFINITION: If the adjective has positive value judgment, and can be removed or replaced without loss in meaning, it is potentially hype.

ADJECTIVE: revolutionary

SENTENCE: This research is innovative because it is the first to examine a revolutionary hypothesis on the origin of the second zygotic centriole, using a mammalian model specifically designed for this purpose.

Figure 5: Prompt fed to LLAMA3.1-INST. and GPT-4O-MINI with a broad definition of hype.

Model	Batch	WD	LR
BERT❄	32	0.00	2e-05
+Finetuning	8	0.10	2e-05
DISTILBERT❄	16	0.01	3e-05
+Finetuning	16	0.00	2e-05
BIOMEDBERT❄	16	0.00	2e-05
+Finetuning	32	0.01	2e-05
GPT-2❄	8	0.01	3e-05
+Finetuning	32	0.01	2e-05

Table 6: Hyperparameters for finetuning the PLMs. Batch refers to the training batch size, WD refers to weight decay and LR refers to learning rate.

Strict Definition Prompt

You are an expert in linguistics, science communication and biomedicine. Following the DEFINITION of hype, determine whether the ADJECTIVE is used in promotional manner in a given SENTENCE. Output HYPE if yes, NOT HYPE if not. Output only the decision. Do not output reasoning.

DEFINITION: If the adjective has positive value judgment, and can be removed or replaced without loss in meaning, it is potentially hype. If the adjective can be removed without loss of meaning, it is considered HYPE.

ADJECTIVE: revolutionary

SENTENCE: This research is innovative because it is the first to examine a revolutionary hypothesis on the origin of the second zygotic centriole, using a mammalian model specifically designed for this purpose.

Figure 6: Prompt fed to LLAMA3.1-INST. GPT-4O-MINI with a stricter definition of hype.

Adjective	Hyperbolic	Gratuitous	Amplified	Coordinated	Broader context
creative	0	33	2	11	19
emerging	1	22	0	6	9
first	0	1	0	0	6
groundbreaking	50	3	0	8	12
innovative	1	40	3	6	17
latest	1	26	0	3	12
novel	1	17	0	4	6
revolutionary	50	2	1	3	18
unique	1	31	1	5	14
unparalleled	50	0	0	3	16
unprecedented	50	5	0	0	16
accurate	0	36	3	10	4
advanced	0	29	0	0	1
careful	0	49	0	1	1
cohesive	0	41	1	7	7
detailed	0	43	3	3	3
nuanced	0	41	1	4	1
powerful	3	42	1	8	1
quality	0	16	0	1	5
reproducible	0	24	2	19	5
rigorous	0	45	1	12	9
robust	0	25	0	7	8
scientific	0	18	1	5	7
sophisticated	0	37	2	5	9
strong	0	31	0	2	4
systematic	0	26	0	5	10
Total	208	683	22	138	220

Table 7: Distribution of rationales for 'hype' classification following the annotation guidelines.

Adjective	%
creative	79.17
emerging	80.00
first	84.00
groundbreaking	86.00
innovative	86.00
latest	70.21
novel	83.67
revolutionary	76.00
unique	72.92
unparalleled	76.00
unprecedented	84.00
Average	79.81

Table 8: Percentage of samples in the training set - NOVELTY.

Adjective	%
accurate	92.00
advanced	87.23
careful	87.76
cohesive	70.00
detailed	80.00
nuanced	86.00
powerful	77.55
quality	84.44
reproducible	81.63
rigorous	83.33
robust	77.55
scientific	74.51
sophisticated	72.92
strong	71.43
systematic	73.47
Average	79.98

Table 9: Percentage of samples in the training set - RIGOUR.