Best-Worst Scaling of Hype in Biomedical Research: Building an Intensity Lexicon of Promotional Adjectives

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

Abstract

Promotional language, or "hype", is increasingly common in biomedical research reporting. Adjectives such as *groundbreaking*, *robust*, and *impactful* can engage readers but also risk imposing value judgements and undermining objectivity. Detecting and assessing such language requires distinguishing degrees of promotional intensity (e.g., *new* < *novel* < *groundbreaking* < *revolutionary*), yet no such graded resource exists. We present an intensity-scaled lexicon of 303 promotional adjectives attested in biomedical writing across eight evaluative domains (e.g. IMPORTANCE, NOVELTY, RIGOUR). Ratings were obtained through Best–Worst Scaling (BWS) with human participants evaluating adjectives for promotional strength in the context of scientific research reporting. We refer to this as the HYPLEX resource (Hype Lexicon). The ratings show high internal consistency (r = 0.87; 95% CI [0.85, 0.89]) and correlate most strongly with arousal and dominance in the NRC VAD Lexicon, suggesting that promotional intensity aligns more with reader activation and perceptions of assertiveness than simple positivity. We also release an online BWS platform integrated with the R package `bwsTools` to support intensity-scaling research in other domains here.

Keywords: promotional language, best-worst scaling, lexical resources, corpus creation

1. Introduction

Although scientific writing is typically characterised as factual, neutral, and objective, authors may also select to use language that promotes a favourable evaluation of their work. For example, importance may be described in absolute terms (*imperative*), novelty sensationalised (*revolutionary*), scale amplified (*vast*), or problems dramatised (*devastating*). Such language has been termed "hype" and defined as "hyperbolic or subjective language employed to glamorise, promote, or exaggerate aspects of research" (Millar et al., 2019).

Our work is motivated by concern about increasing levels of hype in biomedical communication. Promotional language has, for example, risen sharply across the National Institutes of Health (NIH) funding ecosystem, from calls for grants, to applications, and subsequent publications (Millar et al., 2022a,b, 2023). Comparable trends are seen in research articles in other fields, press releases and media reports (Sumner et al., 2014; Weidmann et al., 2018).

As a former editor-in-chief of the JAMA Network journals has pointed out (Bauchner and Rivara, 2022), promotional words such as *groundbreaking* or *transformative* are rarely justified and risk undermining objective assessment, thereby impeding the development of further studies, policies, clinical practice, and knowledge translation. More broadly, public trust in science is weakened when promotional language creates unrealistic expectations or misrepresents findings. We believe that these issues point to the need for systems that assess the promotionality of scientific texts and provide feedback to stakeholders.

The automatic identification of hype poses some challenges for an NLP approach. Firstly, promotionality is often context-dependent. While some terms are used almost invariably with a promotional connotation (e.g. *unprecedented*), others may also carry technical meanings (e.g. *meticulous care was taken when...* vs. *meticulous hemostasis*) or neutral meanings (e.g. *this is the first study to...* vs. *we first analysed...*). Secondly, judgments about whether a term is promotional are inherently subjective. This makes it difficult to establish a "ground truth" for training and evaluation. Finally, systems to assess the promotionality of scientific text require a way to measure the intensity of individual terms. This is because hype should be seen not as a binary distinction but a continuum of intensity, with terms conveying different levels of promotion (e.g. *new* < *novel* < *groundbreaking* < *revolutionary*). Here, we focus on this final challenge.

In this paper, we describe how we obtained promotional intensity ratings for 303 adjectives that are common in biomedical texts and often carry a promotional meaning. The selection of terms is based on prior corpus analyses of promotional language in NIH grant abstracts (Millar et al., 2022a). The adjectives fall into eight semantic categories: IMPORTANCE (e.g., *imperative, paramount*), NOVELTY (*revolutionary, ground-breaking*), SCALE (*massive, vast*), RIGOUR (*careful, sophisticated*), UTILITY of expected outcomes (*impactful, seamless*), QUALI-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8317–8326
11–16 May 2026.

TIES of investigators and environments (*renowned, stellar*), ATTITUDE of the researcher (*incredible, excellent*), and PROBLEM (*dire, devastating*).

We apply Best–Worst Scaling (BWS) to overcome limitations of traditional rating scales. We present intensity ratings for each category and show high inter-rater consistency (split-half reliability = 0.87). Our main contribution is the HYPLEX resource, a lexicon of 303 hype adjectives in biomedical writing with human-derived promotional intensity scores. We compare these scores across semantic categories and validate them against external affective resources. The HYPLEX lexicon is released as supplementary material, and we also provide a web-based BWS annotation tool that integrates with the R package `bwsTools` here.

2. Related Work

2.1. Promotional Language

In linguistics, promotion is usually discussed as part of the broader study of evaluative meaning. Relevant frameworks include Hunston and Thompson (2000)'s model of evaluation, Martin and White (2003)'s appraisal framework, and Hyland (2005)'s stance and engagement framework. All describe systems or resources that enable writers to take positions and attribute value in texts - for example, markers of attitude (*important, valuable*), epistemic stance (*likely, possibly*), degree (*highly effective* vs. *somewhat effective*), and reader alignment (*notably, as we know*). Promotion can be understood as the use of such resources with the intent of encouraging a favourable evaluation.

Evaluative meaning is inherently context dependent, and therefore linguistic analyses tend to rely on manual interpretation to determine if and how expressions convey evaluation. Automatic NLP approaches, in contrast, have typically abstracted away from context, operationalising evaluation as sentiment (positive, neutral, or negative). Recent NLP work (Batalo et al., 2026) has addressed promotional language directly by formalising hype detection in NIH grant texts and using annotated data to develop supervised classifiers for identifying promotional expressions.

2.2. Related Lexicons

Work in sentiment analysis has produced several general-purpose resources for modelling subjectivity and evaluation. The MPQA Subjectivity Lexicon (Wilson et al., 2009) was created through human annotation, with 8,000 plus entries tagged for polarity and subjectivity strength. Connotation Lexicons (Feng et al., 2013) have been developed using semi-automatic induction algorithms to capture implied positive/negative associations, perspective, and value judgments.

Affective lexicons with interval-scaled scores have been constructed using Best-Worst Scaling (BWS). The NRC Valence, Arousal, and Dominance (VAD) Lexicon (Mohammad, 2018) provides ratings for some 20,000 English words. 'Valence' reflects the degree of positivity or negativity, 'arousal' indicates the level of activation or energy, and 'dominance' represents the perceived sense of control or power associated with a word. The NRC Emotion Intensity Lexicon (Mohammad and Bravo-Marquez, 2017) captures graded associations with basic emotions - anger, fear, joy, and sadness. These lexicons are both built for broad-domain text and are not specific to scientific or promotional language.

In the biomedical field, Millar et al. (2022a) identified 139 adjectives that often carry promotional meaning through corpus analyses of successful NIH grant application abstracts. These adjectives showed sharp increases in frequency in both NIH proposals and PubMed abstracts over recent decades. The lexicon has been used in further analyses showing that the proportion of promotional language in proposals is strongly associated with funding success, innovativeness, and subsequent citation impact (Peng et al., 2024; Qiu et al., 2024). However, this resource is limited to 139 adjectives that shifted in frequency and does not quantify promotional intensity (e.g., *novel* < *revolutionary*).

2.3. Best-Worst Scaling

Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991) is a comparative judgment method in which respondents are shown a small set of items and asked to select the one that best and the one that least matches a target property. Because choices are relative rather than absolute, BWS reduces biases common in rating scales (such as middle-ticking or personal scale drift) and produces more consistent and discriminating results across individuals (Finn and Louviere, 1992; Flynn and Marley, 2014).

Each best-worst judgment implicitly generates a series of pairwise preference statements. For example, when presented with four items, if A is judged best and D worst, this implies $A > B, A > C, A > D, B > D$, and $C > D$. Thus, a four-item choice implies five pairwise preferences. When aggregated across many judgments and participants, these comparative data can produce stable estimates of each item's relative position on an underlying scale (Louviere et al., 2015).

BWS has been applied in fields such as marketing, psychology, and linguistics to measure preferences, attitudes, and word meanings (Crocker and Thomson, 2014); (Kiritchenko and Mohammad,

Category	Examples	#
ATTITUDE	outstanding, impressive	20
IMPORTANCE	critical, essential	29
NOVELTY	novel, groundbreaking	31
PROBLEM	devastating, stark	30
QUALITIES	talented, cohesive	44
RIGOUR	systematic, robust	44
SCALE	large-scale, extensive	48
UTILITY	effective, transformative	57
	Total	**303**

Table 1: Sematic categories, example adjectives and number of items in HYPLEX (minus low and neutral anchors).

2016). It has been used to create intensity lexicons for affective variables such as valence, arousal, dominance, and emotion categories (Mohammad, 2018; Mohammad and Bravo-Marquez, 2017).

In this study, we apply BWS to promotional adjectives in biomedical writing in order to generate an intensity scale that quantifies how strongly each term conveys promotional meaning.

3. Methods

3.1. Selection of terms

We limit our focus to adjectives, as this class is most closely associated with evaluation (Martin and White, 2003). Seed terms were drawn from Millar et al. (2022a), which identified 139 "hype" adjectives through longitudinal analysis of corpus of 901,717 abstracts from successful NIH funding applications (NIH corpus). These terms are grouped into eight semantic categories reflecting typical targets of promotional intent.

To expand coverage, we used WordNet (Miller, 1995) and ChatGPT[1] to generate near-synonyms for each seed adjective. Generated terms were checked against the NIH corpus, a collection of 901,717 abstracts from successful U.S. National Institutes of Health (NIH) grant applications spanning 1985-2020 (Millar et al., 2022a), and were retained if they occurred more than ten times and conveyed promotional meaning. The combined final list contained 303 adjectives, distributed across categories as shown in Table 1.

3.2. 3.2 Annotation via Best-Worst Scaling

Participants. Ten annotators took part, consisting of six postgraduate students, two researchers and two university Professors. All were expert users of English, including four native speakers.

[1]https://openai.com/index/introducing-gpt-5/

Anchor Type	Examples	Interpretation
High ($\approx$ 1.00)	revolutionary, flawless	Defines upper bound for category.
Neutral ($>$ Low)	standard, adequate	Quality-check; not used in scaling.
Low ($\approx$ 0.00)	unoriginal, unreliable	Defines lower bound for category.

Table 2: Anchor types, examples, and interpretation. Anchors relevant to each category were embedded within the BIBDs to provide consistent reference points for cross-BIBD scale calibration.

Design. Annotation followed a Balanced Incomplete Block Design (BIBD) (Cochran and Cox, 1957) that was constructed separately for each semantic category (e.g., RIGOUR, NOVELTY, IMPORTANCE). Within each category-specific BIBD, adjectives were presented in sets of four, and annotators selected the most promotional ("best") and least promotional ("worst") term in each set. Anchors relevant to the category were embedded within the BIBDs to provide consistent reference points for cross-BIBD scale calibration. High anchors defined the upper bound of promotional intensity, low anchors defined the lower bound, and neutral anchors served as quality-check items expected to fall directly above the lower bound. Table 2 summarizes the role and interpretation of the anchors with examples for NOVELTY and RIGOUR.

Implementation. BIBDs were created using the `bwsTools` package in R (White II, 2021) and divided into surveys by semantic category. Within each category, all 10 annotators completed the same BWS questionnaire (i.e., identical block sets).

Surveys were administered over five separate days using a custom Best–Worst Scaling (BWS) platform developed for this project, which will be made available upon acceptance. The platform integrates with the R package bwsTools to import balanced incomplete block designs (BIBDs) and automatically generate BWS question sets. The resulting data can then be exported back into bwsTools for statistical analysis. The system was built on a standard web stack and designed to be modifiable for related research applications. A public version of the platform is available at `https://www.hype-busters.com/bws/`.

Each survey began with practice items (unrelated to promotional language), an introduction to promotional language in science, and a category-specific explanation. Attention checks (minimum two, maximum four per survey) were inserted to ensure participants were not responding mechanically.

Figure 1 shows the category instructions for AT-

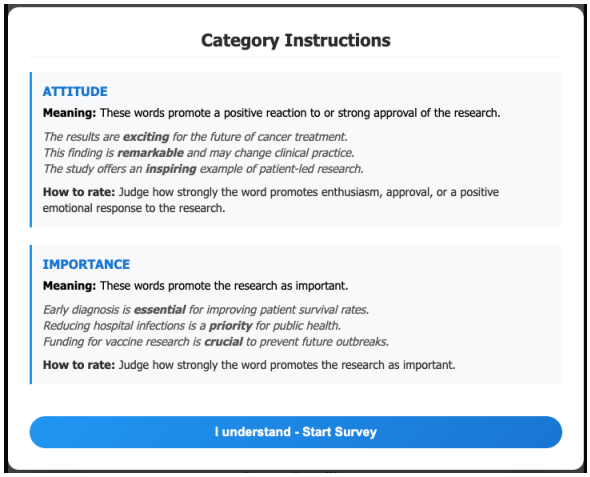

Figure 1: Category instructions for ATTITUDE and IMPORTANCE. For each section, instructions for each category were given, explaining the meaning of each category, as well as giving examples and directions on how to rate the adjectives.

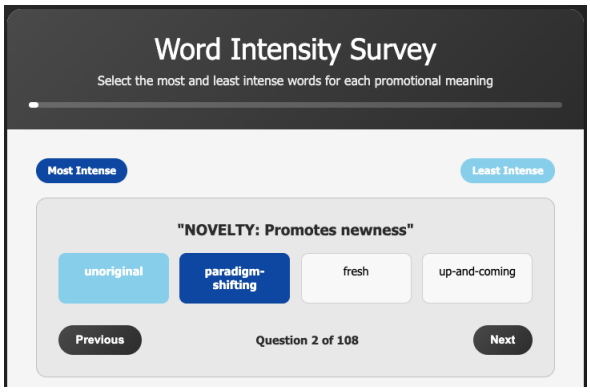

Figure 2: Annotation interface. Annotators were asked to select the least intense (light blue) and most intense (dark blue) among four options.

TITUDE and IMPORTANCE and Figure 2 shows the annotation interface for a sample.

Scoring and aggregation. The final promotional intensity scores were calculated using a difference-scoring procedure (Louviere et al., 2013), implemented using the `diffscoring()` function in the `bwsTools` package in R. For each participant, the number of times an adjective was selected as "least promotional" was subtracted from the number of times it was selected as "most promotional" and the resulting difference was standardized by the total number of times the item appeared.

Within each semantic category, participant-level difference scores were first rescaled to a 0–1 range for interpretability, producing participant-level intensity profiles for each BIBD. These scores were then linearly adjusted using the observed mean scores of the embedded high and low anchors as reference points, setting the low anchor to 0 and the high anchor to 1. This anchor-based calibration preserved within-category rank order while approximately aligning category-specific BIBDs to a comparable intensity range.

Category	Low	Neutral	High
ATTITUDE	0.05	0.09	0.84
IMPORTANCE	0.00	0.11	0.86
NOVELTY	0.07	0.09	0.88
PROBLEM	0.06	0.11	0.93
QUALITIES	0.03	0.16	0.82
RIGOUR	0.05	0.20	0.92
SCALE	0.06	0.14	0.78
UTILITY	0.10	0.17	0.88

Table 3: Mean intensity scores (0-1) for low, neutral, and high anchors by semantic category.

The calibrated participant-level scores were averaged across annotators to obtain the final intensity estimate for each adjective. Louviere et al. (2013) argue that averaged difference scores provide reliable estimates of items' latent scale positions with results comparable to more complex discrete-choice models.

The resulting resource, HYPLEX, is a lexicon of 303 adjectives with category-specific intensity scores.

3.3. Reliability and external comparison

Internal reliability

Split-half reliability (SHR) was used to assess the internal consistency of the annotations. For each semantic category, annotators were randomly divided into two groups, and mean adjective scores were independently computed for each half using the 0-1-scaled difference scores. The correlation between the two halves provided an estimate of how reliably annotators ranked items within each category. The process was repeated 500 times to obtain mean and confidence interval estimates. The same procedure was also applied across all items to evaluate reliability at the full-lexicon level. We do not report kappa-style inter-annotator agreement because BWS yields comparative judgments that are aggregated into continuous item scores rather than independent categorical labels.

Anchor validation. To confirm that the anchors operated as intended, we examined the mean intensity scores for each semantic category to verify that (i) low anchors consistently received the lowest scores, (ii) high anchors consistently received the highest scores, and (iii) neutral anchors were positioned just above the low anchors but below the other items.

Category	SHR (Mean r)	95% CI
ATTITUDE	0.94	[0.91, 0.97]
NOVELTY	0.92	[0.87, 0.96]
PROBLEM	0.92	[0.87, 0.96]
QUALITIES	0.89	[0.84, 0.93]
SCALE	0.88	[0.81, 0.93]
IMPORTANCE	0.87	[0.81, 0.92]
UTILITY	0.78	[0.71, 0.85]
RIGOUR	0.75	[0.63, 0.85]

Table 4: Split-half reliability (SHR) of promotional intensity judgments by semantic category.

External comparison. To assess construct validity, we compared the promotional intensity scores in HYPLEX with affective ratings from the NRC Valence–Arousal–Dominance (VAD) Lexicon (Mohammad, 2018). The VAD model captures three psychological dimensions of affective meaning: valence (positivity), arousal (emotional activation), and dominance (sense of control). We follow the rationale of Qiu et al. (2024), who validated the promotional lexicon (Millar et al., 2022a) upon which our list is based. They show that the original 139 promotional words tend to have higher valence and arousal scores than neutral synonyms. We tested whether promotional intensity co-varied with the VAD affective dimensions across individual adjectives.

4. Results

4.1. Data Overview

The ten annotators completed the eight category-specific surveys, generating a total of 4,340 best-worst judgments (434 per participant). All attention-check questions were answered correctly, suggesting adequate engagement with the task. The best–worst judgments were first converted into participant-level difference scores within each BIBD. Within each category, these were then calibrated using the embedded high and low anchors to place BIBDs on a common scale before being aggregated across participants to generate promotional intensity scores for the 303 adjectives.

4.2. Anchor validation

Table 3 summarizes the mean intensity scores of the low, neutral, and high anchors across categories prior to calibration. In every category, the anchors appeared in the expected order (low < neutral < high). Low and neutral anchors consistently ranked as the least promotional items, while high anchors tended to appear at or near the top of their respective distributions. Thus, the anchors generally functioned as intended and were used as reference points for the calibration of BIBDs within individual semantic categories.

Rank	Adjective	Score
1	extraordinary	0.989 ± 0.035
2	superb	0.975 ± 0.045
3	incredible	0.880 ± 0.128
4	phenomenal	0.875 ± 0.287
5	astonishing	0.843 ± 0.305
6	outstanding	0.811 ± 0.130
7	fascinating	0.804 ± 0.156
8	thrilling	0.720 ± 0.134
9	impressive	0.698 ± 0.164
10	exciting	0.646 ± 0.150
11	remarkable	0.635 ± 0.181
12	excellent	0.587 ± 0.050
13	rewarding	0.540 ± 0.051
14	surprising	0.455 ± 0.153
15	attractive	0.413 ± 0.075
16	confident	0.392 ± 0.223
17	notable	0.387 ± 0.202
18	interesting	0.381 ± 0.067
19	appealing	0.360 ± 0.086
20	intriguing	0.339 ± 0.109

Table 5: Example of category level distribution: mean promotional intensity scores and standard deviations for ATTITUDE adjectives.

4.3. Reliability

As shown in Table 4, split-half reliability (SHR) across categories ranged from r = 0.75 (RIGOUR) to r = 0.94 (ATTITUDE). The more reliable categories (e.g., ATTITUDE, NOVELTY) contain adjectives with clearer evaluative meanings (e.g. *excellent*, *groundbreaking*), whereas UTILITY and RIGOUR involve more technical, context-dependent terms (e.g., *effective*, *accurate*, *precise*) which are likely to be interpreted differently by individual annotators. The overall SHR across all 303 adjectives was r=0.87, 95% CI [0.85, 0.89], consistent with accepted thresholds for good internal consistency for psychometric scales ($\geq$ 0.80) (Boateng et al., 2018). These values are also comparable to those reported for other best-worst scaling datasets, such as Mohammad (2018), whose Valence-Arousal-Dominance lexicon achieved correlations around 0.90-0.95 with a substantially larger pool of annotators over 20,000 words.

4.4. Distribution of promotional intensity scores

Figure 3 shows the distribution of promotional intensity scores across categories. Excluding low and neutral anchors, the mean of promotional intensity scores ranged from 0.55 (PROBLEM) to 0.64

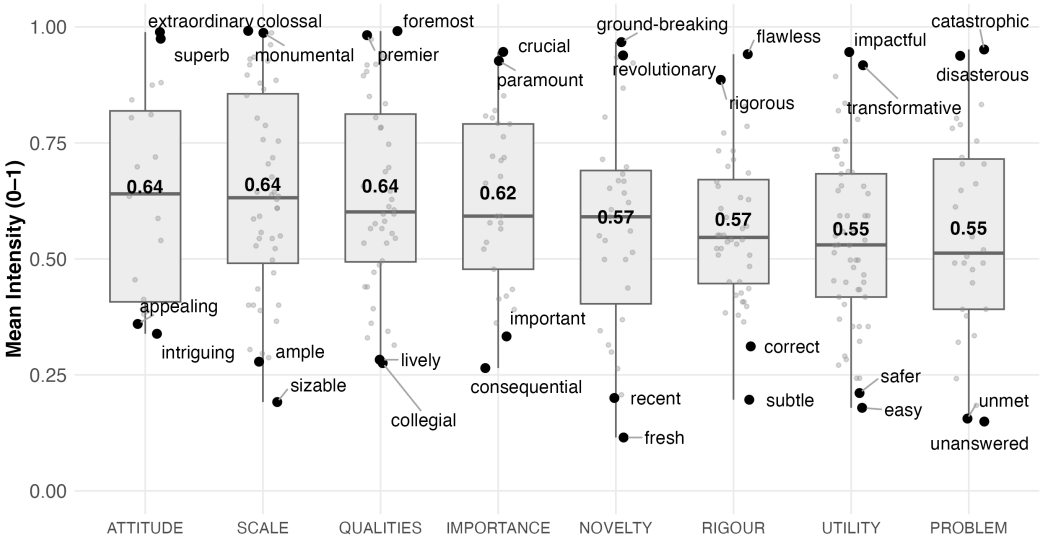

Figure 3: Distribution of Promotional Intensity Scores Across HYPLEX Categories (mean intensity scores shown on boxplots).

(SCALE). Categories showed broadly similar distributions, with interquartile ranges spanning roughly 0.35-0.80. Categories SCALE, QUALITIES, and ATTITUDE included a slightly higher proportion of strongly promotional terms. NOVELTY showed the widest spread of scores (0.12-0.97).

Futhermore, Figure 3 shows the two most and least promotional adjectives in each category. In all cases, these extremes align with intuitive expectations — for instance, *catastrophic* and *disastrous* rank top in PROBLEM, while *unmet* and *unanswered* rank bottom. For illustrative purposes, Table 5 shows the range of scores for the ATTITUDE category.

Finally, Table 6 shows the ranking of all adjectives per category, according to their promotional intensity score. The full set of promotional intensity scores is provided as supplementary material.

	Arousal	Dominance	Valence
UTILITY	0.36*	0.40**	0.26
SCALE	0.57***	0.30*	-0.17
RIGOUR	0.29	0.34*	-0.03
QUALITIES	0.16	0.46**	0.33*
PROBLEM	0.30	0.09	-0.28
NOVELTY	0.55*	0.48*	0.06
IMPORTANCE	0.09	0.16	0.11
ATTITUDE	0.74***	0.78***	0.60**

Figure 4: Correlations Between Promotional Intensity Scores and NRC Valence-Arousal-Dominance (VAD) Dimensions (*$p < .05$; **$p < .01$; ***$p < .001$).

4.5. External comparisons

To contextualize these comparisons, we compare our intensity scores against the NRC VAD Lexicon (Mohammad, 2018). Correlations between our BWS intensity scores and those in the NRC Valence–Arousal–Dominance (VAD) Lexicon were low for valence ($r = 0.11$, $p = .06$) and moderate but significant for arousal ($r = 0.37$, $p < .001$) and dominance ($r = 0.36$, $p < .001$; $n = 280$ overlapping terms).

Figure 4 shows the category-level correlations. ATTITUDE adjectives showed the strongest alignment with VAD ($\approx$ 0.6-0.8 across dimensions), reflecting the general evaluative meanings of adjectives in this category (e.g., *excellent*, *outstanding*, *superb*). In contrast, the other categories, which are more domain specific, showed weaker associations, likely because their adjectives function differently across contexts. Terms such as *thorough*, *exact*, and *precise*, for example, may convey evaluative force in scientific writing but are relatively neutral in general English.

Interpretive context plays a role in how evaluative language is understood. In our study, adjectives were judged within a scientific frame, whereas the VAD lexicon relies on context-free ratings and are likely to reflect more general associations. As such, words like *grave* are likely interpreted as nouns, and *sophisticated* as a social rather than methodological descriptor. Both adjectives are examples of outliers that weaken the observed correlations. In addition, several strongly promotional terms

ATTITUDE	IMPORTANCE	NOVELTY	PROBLEM	QUALITIES	RIGOUR	SCALE	UTILITY
1 extraordinary	1 crucial	1 ground-breaking	1 catastrophic	1 foremost	1 flawless	1 colossal	1 impactful
2 superb	2 paramount	2 revolutionary	2 disasterous	2 premier	2 rigorous	2 monumental	2 transformative
3 incredible	3 critical	3 paradigm-shifting	3 devastating	3 leading	3 meticulous	3 staggering	3 high-performance
4 phenomenal	4 indispensable	4 unprecedented	4 ruinous	4 brilliant	4 robust	4 enormous	4 high-yielding
5 astonishing	5 pivotal	5 unheard-of	5 hopeless	5 prestigious	5 high-level	5 innumerable	5 perfect
6 outstanding	6 imperative	6 unparalleled	6 deadly	6 distinguished	6 high-standard	6 mammoth	6 ideal
7 fascinating	7 vital	7 one-of-a-kind	7 dire	7 exceptional	7 sophisticated	7 overpowering	7 optimal
8 thrilling	8 ultimate	8 unequaled	8 grave	8 stellar	8 painstaking	8 gigantic	8 synergistic
9 impressive	9 momentous	9 unique	9 miserable	9 pre-eminent	9 precise	9 tremendous	9 high-performing
10 exciting	10 essential	10 trailblazing	10 perilous	10 veteran	10 methodical	10 massive	10 meaningful
11 remarkable	11 invaluable	11 unrivaled	11 grim	11 renowned	11 exacting	11 overwhelming	11 valuable
12 excellent	12 fundamental	12 never-before-seen	12 bleak	12 accomplished	12 thorough	12 greatest	12 effective
13 rewarding	13 urgent	13 pioneering	13 desperate	13 thriving	13 accurate	13 vast	13 durable
14 surprising	14 foundational	14 game-changing	14 shocking	14 talented	14 scientific	14 immense	14 efficient
15 attractive	15 pressing	15 cutting-edge	15 dismal	15 gifted	15 error-free	15 sweeping	15 scalable
16 confident	16 profound	16 inventive	16 frightening	16 seasoned	16 powerful	16 worldwide	16 seamless
17 notable	17 high-priority	17 incomparable	17 alarming	17 forward-thinking	17 fine-grained	17 myriad	17 opportune
18 interesting	18 prime	18 innovative	18 daunting	18 reputable	18 strict	18 countless	18 efficacious
19 appealing	19 decisive	19 radical	19 stark	19 longstanding	19 detailed	19 exhaustive	19 concrete
20 intriguing	20 significant	20 latest	20 formidable	20 established	20 careful	20 comprehensive	20 expandable
	21 key	21 emerging	21 intimidating	21 dedicated	21 disciplined	21 global	21 productive
	22 necessary	22 novel	22 scarce	22 credentialed	22 advanced	22 largest	22 sustainable
	23 major	23 first	23 disturbing	23 certified	23 discriminating	23 transdisciplinary	23 useful
	24 chief	24 up-to-date	24 serious	24 successful	24 strong	24 biggest	24 purposeful
	25 compelling	25 original	25 dramatic	25 dynamic	25 elegant	25 far-reaching	25 streamlined
	26 strategic	26 creative	26 worrying	26 skilled	26 structured	26 fastest	26 extensible
	27 influential	27 newest	27 troubling	27 knowledgeable	27 exact	27 generous	27 tailored
	28 important	28 imaginative	28 elusive	28 qualified	28 systematic	28 multifarious	28 intuitive
	29 consequential	29 up-and-coming	29 unmet	29 senior	29 quality	29 huge	29 fruitful
		30 recent	30 unanswered	30 experienced	30 coordinated	30 top	30 dependable
		31 fresh		31 rising	31 reproducible	31 diverse	31 tactical
				32 vibrant	32 cohesive	32 intense	32 straightforward
				33 ambitious	33 verifiable	33 extensive	33 constructive
				34 promising	34 complex	34 expansive	34 accessible
				35 energetic	35 integrated	35 international	35 adaptable
				36 holistic	36 refined	36 abundant	36 actionable
				37 intellectual	37 logical	37 complete	37 implementable
				38 supportive	38 unified	38 considerable	38 practical
				39 motivated	39 empirical	39 deeper	39 generalizable
				40 integrative	40 repeatable	40 immediate	40 practicable
				41 committed	41 nuanced	41 wide-ranging	41 tangible
				42 cohessive	42 organized	42 substantial	42 transferable
				43 lively	43 correct	43 broad	43 maintainable
				44 collegial	44 subtle	44 plenty	44 well-timed
						45 prompt	45 timely
						46 instant	46 rich
						47 ample	47 easy-to-use
						48 sizable	48 user-friendly
							49 deployable
							50 self-explanatory
							51 usable
							52 viable
							53 ready
							54 simple
							55 economical
							56 safer
							57 easy

Table 6: Category-wise ranking of adjectives in the HYPLEX lexicon, ordered by mean promotional intensity score.

(e.g., *groundbreaking*, *paradigm-shifting*, *trailblazing*) were absent from the VAD lexicon.

In sum, differences in coverage, lexical scope, and interpretive context likely account for the generally modest correlations. Nevertheless, the strongest associations were observed for arousal and dominance dimensions, linked to activation and agency in language use, rather than valence, which reflects simple positivity. This seems intuitive if we consider promotional language as aiming to elicit engagement and convey persuasive force.

5. Applications and future work

The HYPLEX resource provides empirically scaled intensity scores for over 300 promotional adjectives attested in biomedical writing. Potential applications and extensions include:

- **Features for NLP models:** Supplying interpretable features for hype-detection systems and for developing hype-aware word and sentence embeddings.
- **Benchmarking:** Serving as a gold-standard reference for evaluating automatic methods of determining promotionality.
- **Diachronic and cross-domain analyses:** Enabling studies that track changes in promotional language across time, research domains, or publication types.
- **Annotator variation:** Supporting analyses of how promotional intensity judgments differ

across groups defined by, for example, linguistic background or expertise (e.g., novice vs. expert, native vs. non-native).

- **Extending resource coverage:** Incorporating other parts of speech (e.g. adverbs, verbs), epistemic stance markers that modify the perceived strength of promotion (e.g. *highly*, *somewhat*) or certainty of claims (*demonstrates*, *suggests*), and dispromotional adjectives (*sloppy*, *inadequate*).
- **Perception and research impact:** Examining whether levels of promotionality influence readers' evaluation of research, or measurable outcomes such as funding success, publication, and citation impact.
- **Resource growth:** Using the online BWS platform to crowdsource additional annotations, refine scores, and enhance reliability.

6. Conclusions

To the best of our knowledge, this is the first attempt to provide empirically derived intensity scores for promotional language in scientific writing. Building on prior analyses of NIH funding applications, we developed the HYPLEX resource - an intensity-scaled lexicon of 303 promotional adjectives attested in biomedical research writing. The lexicon showed overall internal consistency within preferred thresholds for psychometric scale quality and produced intuitively ordered rankings (e.g., *important* < *vital* < *critical* < *paramount*). Quantitative comparisons with affective norms further indicated that promotional intensity aligns most strongly with arousal and dominance dimensions rather than with valence alone. This pattern suggests that what we measure as promotionality primarily captures assertive or attention-directing emphasis rather than simple positivity. These properties make HYPLEX potentially useful for developing systems that detect and assess promotional framing ("hype") in research communication.

Limitations

Our work has limitations:

- **Domain coverage:** The resource is limited to adjectives attested in biomedical research writing. Promotional language may vary across other domains and genres.
- **Linguistics scope:** Although promotion can be realised through a range of linguistic resources, our resource is restricted to adjectives.
- **Annotator heterogeneity:** Participants were not specialists in biomedicine and varied in their linguistic and disciplinary backgrounds. While this likely introduces variability in judgments, it arguably reflects how diverse audiences interpret evaluative language in research texts.
- **Context dependence:** Best–Worst Scaling captures relative preferences independent of textual context, so pragmatic nuances are not well represented.
- **Calibration limits:** Anchor-based calibration supports consistency within categories but may not ensure comparability across categories or domains.
- **Scaling limitation:** BWS provides relative judgments but not absolute magnitudes of promotional strength.

Ethics Statement

We collected and annotated data for this study in accordance with ethical research standards. All participants provided informed consent prior to participation. Student annotators were compensated for their time at rates consistent to local minimum wage standards. Researchers and Professors generously dedicated their free time to assist us. The data contain no personally identifiable information, and all experiments comply with the terms of service of the data sources.

While our proposed lexicon brings awareness to issues in using promotional language in scientific publishing in biomedical research, one may misuse the findings of this paper to purposely include or accentuate 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).

7. Bibliographical References

Bojan Batalo, Erica K. Shimomoto, Dipesh Satav, and Neil Millar. 2026. Hype or not? formalizing automatic promotional language detection in biomedical research. In *Proceedings of the 19th*

Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.

Howard Bauchner and Frederick P Rivara. 2022. The scientific communication ecosystem: the responsibility of investigators. *The Lancet*, 400(10360):1289–1290.

Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149.

William G Cochran and Gertrude M Cox. 1957. *Experimental designs*. John Wiley & Sons.

Christopher Crocker and David MH Thomson. 2014. Anchored scaling in best–worst experiments: A process for facilitating comparison of conceptual profiles. *Food Quality and Preference*, 33:37–53.

Adam Finn and Jordan J Louviere. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, 11(2):12–25.

Terry N Flynn and Anthony AJ Marley. 2014. Best-worst scaling: theory and methods. In *Handbook of choice modelling*, pages 178–201. Edward Elgar Publishing.

Susan Hunston and Geoffrey Thompson. 2000. *Evaluation in text: Authorial stance and the construction of discourse: Authorial stance and the construction of discourse*. Oxford University Press, UK.

Ken Hyland. 2005. Metadiscourse: Exploring interaction in writing. *Continuum*.

Jordan Louviere, Ian Lings, Towhidul Islam, Siegfried Gudergan, and Terry Flynn. 2013. An introduction to the application of (case 1) best–worst scaling in marketing research. *International journal of research in marketing*, 30(3):292–303.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, working paper.

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, Françoise 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.

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.

Huilian Sophie Qiu, Hao Peng, Henrik Barslund Fosse, Teresa K Woodruff, and Brian Uzzi. 2024. Use of promotional language in grant applications and grant success. *JAMA network open*, 7(12):e2448696–e2448696.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349.

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.

Mark H White II. 2021. bwstools: An r package for case 1 best-worst scaling. *Journal of choice modelling*, 39:100289.

8. Language Resource References

Feng, Song and Kang, Jun Seok and Kuznetsova, Polina and Choi, Yejin. 2013. *Connotation lexicon: A dash of sentiment beneath the surface*

meaning. PID https://www3.cs.stonybrook.edu/~ychoi/connotation/.

Kiritchenko, Svetlana and Mohammad, Saif. 2016. *Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling*. PID https://www.saifmohammad.com/WebPages/BestWorst.html.

Miller, George A. 1995. *WordNet: a lexical database for English*. ACM New York, NY, USA. PID https://wordnet.princeton.edu/.

Mohammad, Saif. 2018. *Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words*. Association for Computational Linguistics. PID https://saifmohammad.com/WebPages/nrc-vad.html.

Mohammad, Saif and Bravo-Marquez, Felipe. 2017. *Emotion Intensities in Tweets*. Association for Computational Linguistics. PID https://www.saifmohammad.com/WebPages/BestWorst.html.

Wilson, Theresa and Wiebe, Janyce and Hoffmann, Paul. 2009. *Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis*. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info .... PID https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.