Personalized Jargon Identification for Enhanced Interdisciplinary Communication

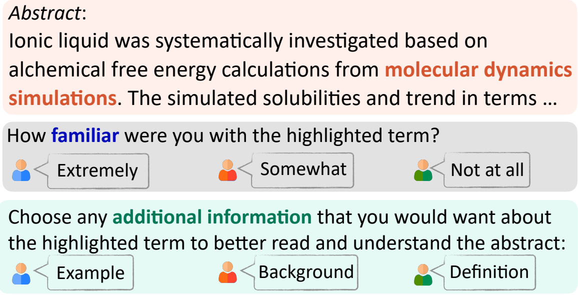
Yue Guo[1,*] Joseph Chee Chang[2] Maria Antoniak[2]
Erin Bransom[2] Trevor Cohen[1] Lucy Lu Wang[1,2] Tal August[2]
[1]University of Washington [2]Allen Institute for AI
{yguo50, cohenta, lucylw}@uw.edu,
{josephc, mariaa, erinbransom, tala}@allenai.org

Abstract

Scientific jargon can confuse researchers when they read materials from other domains. Identifying and translating jargon for individual researchers could speed up research, but current methods of jargon identification mainly use corpus-level familiarity indicators rather than modeling researcher-specific needs, which can vary greatly based on each researcher's background. We collect a dataset of over 10K term familiarity annotations from 11 computer science researchers for terms drawn from 100 paper abstracts. Analysis of this data reveals that jargon familiarity and information needs vary widely across annotators, even within the same sub-domain (e.g., NLP). We investigate features representing domain, subdomain, and individual knowledge to predict individual jargon familiarity. We compare supervised and prompt-based approaches, finding that prompt-based methods using information about the individual researcher (e.g., personal publications, self-defined subfield of research) yield the highest accuracy, though the task remains difficult and supervised approaches have lower false positive rates. This research offers insights into features and methods for the novel task of integrating personal data into scientific jargon identification.[1]

1 Introduction

An important challenge to communicating knowledge across scientific domains is aligning on a shared vocabulary (Strober, 2006). Each scientific domain has unique terminology that optimizes communication within the field but can pose a barrier to researchers in other domains (Lucy et al., 2022; Choi and Pak, 2007). As science becomes more specialized, so too does its terminology (Barnett and Doubleday, 2020; Plaven-Sigray et al., 2017), raising the barrier of learning and collaborating across disciplines. We envision systems that can identify whether specialized terminology will be unfamiliar to an individual scholar, so that other systems can then translate this terminology.



Abstract:
Ionic liquid was systematically investigated based on alchemical free energy calculations from **molecular dynamics simulations**. The simulated solubilities and trend in terms …

How **familiar** were you with the highlighted term?

Choose any **additional information** that you would want about the highlighted term to better read and understand the abstract:

Figure 1: An annotated term from our dataset, with annotations by computer science researchers. Despite sharing a common domain, these researchers exhibit variation in their familiarity and additional information needs about the term within the abstract. Abstract from Liu et al. (2014).

NLP techniques have been developed to identify and simplify scholarly jargon (Gardner and Davies, 2013; Tanaka-Ishii and Terada, 2011; Guo et al., 2022, 2021), a first step in our envisioned setting. The majority of these techniques use a corpus of documents as a proxy for what a reader knows (e.g., Wikipedia is assumed to contain words known to a general audience). However, an individual's specific background knowledge also plays a role in determining their familiarity with a word (Gooding and Tragut, 2022). For example, a theoretical computer science (CS) researcher might struggle more with jargon in a chemistry paper than in a mathematics paper, but the opposite may be true for a computational biologist. Information on a researcher's background should help determine what they already know and what they need explained.

In this paper, we introduce the task of *personalized scholarly jargon identification*. We ground our investigation in the real-world setting of interdisci-

[*] Work performed during internship at AI2.

[1] The code and anonymized version of our dataset are available at: https://github.com/talaugust/PersonalizedJargon.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4535–4550
June 16-21, 2024 ©2024 Association for Computational Linguistics

plinary reading: researchers reading papers in less familiar domains. We first validate our setting with an initial study on interdisciplinary reading. The results reveal a clear preference for supplementary information beyond what is provided in a paper abstract, especially in less familiar domains. Building on initial findings, we propose the task of predicting the familiarity and any associated information needs of a term for an individual researcher.

To study this problem, we collect a dataset of over 10K individual familiarity ratings and information needs from 11 CS researchers about terms drawn from 100 out-of-domain abstracts (example in Figure 1). We enumerate features representing an individual's background knowledge based on papers they have written and read. Using these features and our dataset, we investigate baselines for estimating term familiarity, including regression models and prompt-based approaches using large language models (LLMs). Our analysis reveals that incorporating individual-level information improves the accuracy of predicting term familiarity, though the task is difficult and no one model performs best for all annotators. Our project contributes the following:

- We define the novel task of predicting personalized jargon familiarity. We motivate our task based on initial experiments with interdisciplinary computer science researchers.
- We collect a dataset of over 10K term familiarity ratings and individual information needs.
- We enumerate features representing an individual researcher's knowledge and investigate integrating these features into supervised and prompt-based methods.

2 Related Work

Interdisciplinary communication Interdisciplinary research integrates knowledge from multiple disciplines to address a shared question (Daniel et al., 2022). Choi and Pak (2007) surveyed interdisciplinary researchers in the health sciences, finding that a mismatch in terminology complicates efforts in communicating between disciplines. Lucy et al. (2022) found that papers that used more discipline-specific terminology (i.e., jargon) had fewer citations across disciplines, and Martínez and Mammola (2021) found that papers that use more jargon are generally cited less.

Scientific text simplification Detecting scholarly jargon is commonly done using corpus-based approaches (Tanaka-Ishii and Terada, 2011). For example, Gardner and Davies (2013) identified scholarly jargon in English by studying the frequency of words within scientific papers compared to a background corpus of general English writing. Similar methods have identified jargon in specific fields of science, including medical studies (August et al., 2022), and computer science papers (Salatino et al., 2018). Gooding and Tragut (2022) found that training models at the individual level improves general English complex word identification, and Lin et al. (2012) found that using social media posts written by an individual can help predict word familiarity. Murthy et al. (2021) generated alternate definitions of scientific terms to better align with a scientist's knowledge. Work has also explored interactive systems to augment scientific abstracts (Fok et al., 2023) and provide term definitions (Head et al., 2020; August et al., 2022).

In contrast to prior work, we focus on predicting familiarity of scholarly jargon. Our focus on scientists provides unique opportunities to model individual knowledge. Scientists develop deep knowledge of their field by reading and writing scientific papers. Models that can achieve accurate, individualized predictions for scientists could greatly improve interactive systems by focusing aids (e.g., definitions, additional information) on only the unfamiliar words for an individual reader (Ridder, 2002; Head et al., 2020).

3 Task Description

We conduct an initial study with 10 computer science researchers to (i) validate our intuition that term familiarity is important for interdisciplinary reading, and (ii) identify what information needs researchers have for unfamiliar terms when reading across domains.

We recruited participants from two subdomains of computer science: from Natural Language Processing (NLP) and from Human-Computer Interaction (HCI). Participants were asked to read two paper abstracts, one from a closer domain (Linguistics or Psychology) and one from a more distant domain (Medicine). For each paper, we provided two abstract variants: the original author-written abstract and a generated abstract personalized to the participant's background. To personalize the abstracts, we prompted GPT-3 (*text-davinci-003*) to rewrite the provided abstract as an abstract personalized to an author. The model was given the author's paper abstracts and a sampled list of citation sentences from the author's papers (details can be found in App. Table 7). After reading each abstract pair, participants identified what modifications they liked/disliked in the personalized abstract, and provided a free-text response on whether they preferred the generated abstract and why. The study was considered exempt by University of Washington's IRB.

3.1 Initial Findings

We collected a total of 20 responses from 10 researchers (7 from NLP and 3 from HCI). Participants generally preferred the personalized abstract over the original, with 9 preferring the personalized abstract for the medical paper (90%) and 5 preferring the personalized abstract for the linguistics or psychology paper (50%). There was a general preference for modifying the abstract by *adding* information (82% of additions preferred) over *removing* it (9% of removals preferred). Full results in App. Figure 7.

We categorize participants' preferred modifications as satisfying the following information needs:

- **Definition**: key information about the term independent of any context. A definition answers the question, "What is/are [term]?"
- **Background**: information that is important for understanding the term in the context of the abstract, e.g., how the term relates to the overall problem, significance, and motivation.
- **Example**: specific instances that help illustrate the usage of the term within the abstract.
- **Method/Result Details**: details on the methodology and results of the paper.
- **Relevant Downstream Connections**: insights about how the current paper's findings relate to the reader's own research.

The first three information needs pertain to additional information for specific terms, while the latter two require further contextualization of the information in the abstract. Participants in our study generally requested additional term-specific information when they were *less* familiar with the term and domain, and requested contextualizing information when they were *more* familiar with the domain. Given the more clear association between the need for term-specific information and term/domain unfamiliarity, we focus on the first three needs—definitions, background, and examples—in the remainder of this work. We also note that while participants generally reacted positively to relevant downstream connections in all cases, these texts were usually hallucinated by the model, so we avoid targeting these as well. Examples of modifications that the models made to the abstracts when personalizing are provided in App. Table 6.

3.2 Task Definition

Based on these initial findings, we identify the tasks of individual term familiarity prediction and information need prediction as important steps for assisting interdisciplinary reading of scientific abstracts. We formalize the first task as: given an individual researcher defined by their authored publications $R = \{r_1, r_2, ...r_m\}$ and an abstract to personalize A, which includes a set of terms $T = \{t_1, t_2, ...t_n\}$, our goal is to predict the subset of terms unfamiliar to R. In addition, we aim to predict R's indicated information need from among {definition, background, example} for each term, as defined above.

4 Dataset

As no pre-existing datasets exist for personalized scientific jargon identification, we construct a new dataset of terms from abstracts with human annotations of familiarity and additional information needs. We direct our focus to abstracts that are outside the individual's domain, with CS researchers as the annotators.

4.1 Data Source

To ensure that the out-of-domain abstracts could realistically be read by our annotators, we compile a corpus of non-CS papers often viewed by CS researchers, published after 2010, using the Semantic Scholar API (Kinney et al., 2023). We define CS researchers as anyone who has (co-)authored a paper categorized as 'Computer Science' as classified in the API. From the top 500 viewed papers *not* categorized as CS, we take a stratified sample of 100 abstracts covering the 22 domains found in the top 500 papers (paper counts are in App. Figure 8).

For each abstract, the top-10 significant terms are identified using the OpenAI model *text-davinci-003* (details and prompt in App. Table 7). We manually review 10 abstracts and confirm that their top 10 terms align with our notion of salient terms in each abstract—we define salient terms as terms that could be provided as keywords of the paper.

4.2 Annotation

Annotators were asked to annotate each term with the following information:

- **Familiarity** on a scale of 1 (not at all familiar) to 5 (extremely familiar). Not at all familiar was defined as "you have never heard of this term." Extremely familiar was defined as "you have a deep, comprehensive understanding of this term."
- **Additional information needs** that could help annotators understand the abstract. These included definitions, background, and examples for each term (defined in §3.1). Annotators could select more than one information need per term.

We recruited 11 annotators from UpWork with a Master's (N=4) or Doctorate (N=7) degree in CS who had published at least one paper (see Table 1). We paid each annotator $20-30 hourly based on their degree. Each annotator reviewed all sampled abstracts and answered all questions, providing a total of 10,571 familiarity ratings for 956 terms.[2]

Familiarity data check To ensure that annotator familiarity ratings were consistent, we conducted a data check for all annotators. We selected 10 entities from each annotator, 5 rated as familiar and 5 rated as unfamiliar. For each entity, we asked annotators to provide a definition of the entity without looking up any information. If they could not define the term, we instructed them to write 'N/A'. Annotators were generally consistent with their initial scores, with 81% of responses matching initial ratings (i.e., if they were familiar, they wrote a correct definition). When initial scores did not align with the data check, annotators generally wrote definitions based on the term's context in the abstract.

4.3 Outcomes

We define a binary term familiarity outcome measure by grouping the collected 5-point familiarity ratings into the binary classes of "familiar" (ratings ≥ 3) and "unfamiliar" (ratings ≤ 2). We treat the need for additional definitions, background, and examples as separate binary classification tasks (covered by RQ5 in §6) .

ID	Degree	# of Papapers	Self-Defined Subfield
1	Master	20	Computer Vision
2	PhD	10	Networking
3	Master	1	NLP
4	PhD	20	NLP
5	PhD	30	Cyber Security
6	PhD	4	General CS theory
7	PhD	3	Neural Networks
8	PhD	60	NLP
9	PhD	15	Complex Networks
10	Master	2	Computer Vision
11	Master	2	Computer Vision

Table 1: Annotators' characteristics

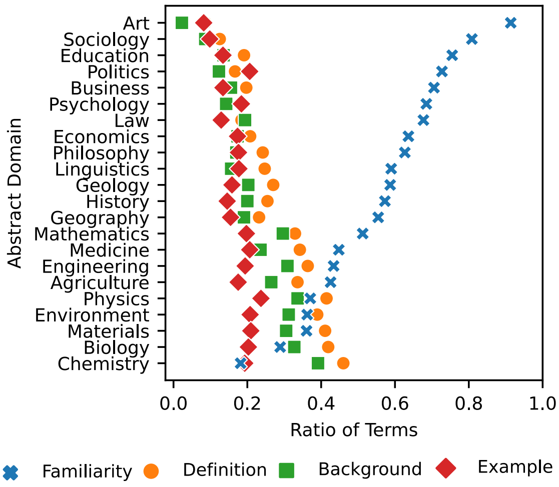


Figure 2: Mean familiarity and additional information needs (definition, background, and example) across abstract domains. The ratio of terms shows how many terms in the abstract domain are familiar, and require definitions, background, and examples.

4.4 Analysis

Below we describe characteristics of our dataset, focusing on how familiarity ratings and information needs exhibit variation across abstract domain and annotator background.

Domain-specific variation Figure 2 illustrates the differences in familiarity and additional information needs across abstract domains. Annotators were most often familiar with terms from Art, while Chemistry received the lowest familiarity ratings. This is in line with prior work, which has suggested that the technical sciences often develop more specialized vocabulary within a domain, while the social sciences and humanities share more terminology between domains (Lucy et al., 2022; Vilhena et al., 2014). Mathematics, which in prior work has been found to contain a large amount of discipline-specific terminology (Lucy et al., 2022), was one of

[2] GPT 3.5 identified <10 terms from some abstracts. Upon inspecting these abstracts, the authors agreed that there were fewer than 10 salient terms to list.

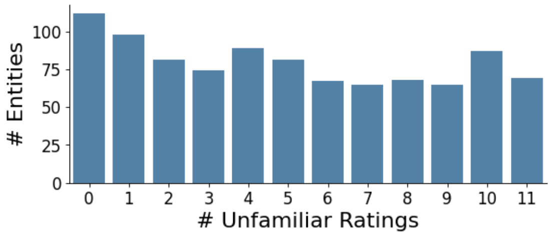


Figure 3: The number of terms rated as unfamiliar by annotators, broken down by the number of annotators. Generally there is a uniform number of ratings for each level of agreement.

the most familiar outside the social sciences. One possible reason was that annotators were generally from CS sub-domains that share overlap with Mathematics (e.g., ML, Computer Vision). Examples of terms and annotations are in App. Table 12.

The same trend we observed for familiarity ratings held for definition and background information requests. However, annotators preferred a roughly constant rate of examples regardless of domain. Looking at common terms that annotators requested examples for, we see that generally these terms refer to a category rather than a single concept. For example, annotators requested examples for terms in the humanities like "mental operations" (5/11 annotators) and "mass communication technologies" (6/11 annotators) even when most annotators rated these terms as familiar (10/11 and 9/11, respectively).

Individual-specific variation There is substantial variation in term familiarity across annotators. As Figure 3 shows, terms vary widely in how many annotators rated them as familiar, with a slightly greater number of terms being rated as familiar by all annotators. For 15% of the terms, half the annotators disagreed (i.e., 5 or 6 deemed it familiar while the rest did not). The field most commonly found in this split was Mathematics (12% of terms).

Annotator backgrounds were associated with what terms they found familiar. Taking Mathematics again as an example, two annotators self-identifying as working in CS Theory and Neural Networks rated Mathematics terms as more familiar (mean=72%, std=4%) compared to other annotators (mean=44%, std=22%). There is similar variation for Linguistics. The three annotators identifying as NLP researchers rated an average of 64% of terms from Linguistics abstracts as familiar with little variation between them (std=6%). The remaining annotators rate an average of 57% Linguistics terms as familiar, with much greater variation (std=20%).

5 Prediction

Given our dataset and annotator backgrounds, we investigate the effectiveness of a set of features and methods for predicting individual term familiarity and additional information needs.

5.1 Features

Past work has explored using readability measures, frequency statistics, and embeddings to predict term familiarity at the population-level (e.g., for all lay readers) (Rakedzon et al., 2017; August et al., 2022; Lucy et al., 2022). We adapt the following features for predicting individual-level familiarity:

- **Frequency**: The number of times a term appears in a researcher's publications $R = \{r_1, r_2, ...\}$.
- **Specificity**: The term's uniqueness to a corpus (Zhang et al., 2017), computed as the log probability ratio:

  $$S_c(t) = log \frac{P_c(t)}{P_C(t)}$$

  In our case, c corresponds to the target abstract A and C to the researcher's publications R.
- **Embedding similarity**: The minimum Euclidean distance between the target abstract A's embedding and any of the author's publications $R = \{r_1, r_2, ...\}$'s embeddings. We use embeddings from SPECTER 2.0 (Singh et al., 2022), a citation-based transformer model encoding semantic document similarity.

For each feature, we start by defining different granularities of a researcher's publications R, representing domain, subdomain, and individual-level information. The following data is extracted from the Semantic Scholar API (Kinney et al., 2023):

- **Domain**: 10K randomly sampled CS papers from 2015-2022.
- **Subdomain**: 10K randomly sampled papers from each annotator's self-defined CS subdomain from 2015-2022. Subdomains are defined manually based on venues associated with a given subdomain.
- **Individual**: All an individual's publications. If the individual's number of publications is less than the necessary number for training in §5.2, the remaining quantity is supplemented by a random selection from the cited references within those publications.

In addition to these granular features, we include the following general-purpose measures of readability and metadata:

- **Readability**: We use the Flesch-Kincaid (F-K) (Flesch, 2007) readability score and the GPT-2 perplexity score (Martinc et al., 2021). F-K score is computed at the passage level; all terms from an abstract are assigned the same F-K score.
- **Metadata**: We include the target paper domain, the year of the annotator's first published paper, total number of published papers, and the annotators average citation count for published papers.

5.2 Models

We explore two modeling approaches: supervised and prompt-based.

Lasso regression We adopt a logistic regression model with L1-regularization to integrate features across various levels of granularity and determine feature importance. A binary label determination is made using a threshold value of 0.5. We train one model per annotator. There were two training settings:

- *Individual model*: the model is trained using data from one annotator to predict ratings from the same annotator.
- *Mixed model*: the model is trained on ratings from all other annotators (i.e., leave-one-annotator-out testing). To maintain the same sample size of training data as the individual Lasso models, we randomly select the same number of training data points as was used to train an annotator's individual model.

Prompt-based LLMs We design prompts to predict binary term familiarity using the GPT-4 model from OpenAI in September 2023. To explore different strategies, we use:

- *Baseline*: providing only the term and abstract containing the term. This is essentially a zero-shot setting with no personalization information.
- *Metadata*: providing annotator metadata along with the baseline prompt.
- *Context-enhanced learning*: providing publications at either the annotator's domain, subdomain, or individual level.
- *Few-shot learning*: providing examples of term-abstract-rating tuples from our labeled dataset. Ratings are drawn from the three levels of granularity: ratings from other annotators with no overlap in subdomain (domain), from other annotators in the same subdomain (subdomain), and from the annotator (individual).

Model	F1	Recall	Precision
Majority Baseline	62.9$_{\pm 1.4}$	100.0$_{\pm 0.0}$	45.9$_{\pm 1.5}$
Oracle			
Majority	71.5$_{\pm 1.7}$	69.6$_{\pm 2.1}$	73.4$_{\pm 2.2}$
Nearest-neighbor	71.9$_{\pm 1.7}$	76.0$_{\pm 2.1}$	68.2$_{\pm 2.1}$
Lasso			
Mixed	56.9$_{\pm 1.9}$	59.5$_{\pm 2.3}$	54.6$_{\pm 2.3}$
Individual	60.6$_{\pm 2.0}$	55.3$_{\pm 2.3}$	67.2$_{\pm 2.3}$
GPT			
Baseline	63.1$_{\pm 1.5}$	100.0$_{\pm 0.0}$	46.1$_{\pm 1.6}$
Metadata	64.0$_{\pm 1.5}$	94.4$_{\pm 1.0}$	48.4$_{\pm 1.6}$
Context-enhanced	64.2$_{\pm 1.5}$	98.7$_{\pm 0.5}$	47.6$_{\pm 1.6}$
Few-shot	62.8$_{\pm 1.5}$	99.6$_{\pm 0.3}$	45.8$_{\pm 1.6}$

Table 2: Mean model performance (±std) on term familiarity prediction in the test set. Standard deviation is estimated by bootstrapping with 1,000 resamples, each made up of 1,000 labels. Context-enhanced and few-shot learning are prompted with individual-level data. Metadata model are prompted with all metadata.

For context-enhanced learning and few-shot learning, we experiment with 1, 5, and 10 examples.

Oracle settings We also include two oracle classifiers based on social recommendations and collaborative filtering (Kang et al., 2022; Guy et al., 2009; Goldberg et al., 1992):

- *Majority oracle*: the majority rating from all other annotators.
- *Nearest-neighbor oracle*: the annotator's rating who has the most similar ratings to the current annotator (i.e., having the highest agreement in ratings). Similarity of ratings is based on the training set. Nearest-neighbor pairings are listed in App. Table 11.

5.3 Evaluation

We split the entities randomly into an 80/20 train/test set split for each annotator, with the test set containing 2200 ratings across 200 entities. All models are evaluated on the same test set per annotator, reporting F1 score, recall, and precision to measure classification performance. To identify critical features in a Lasso model, we count the features with non-zero value coefficients. A higher frequency denotes greater and more consistent influence of the feature on prediction.

6 Results

RQ1. How do supervised and prompt-based methods perform? In Table 2, we present a com-

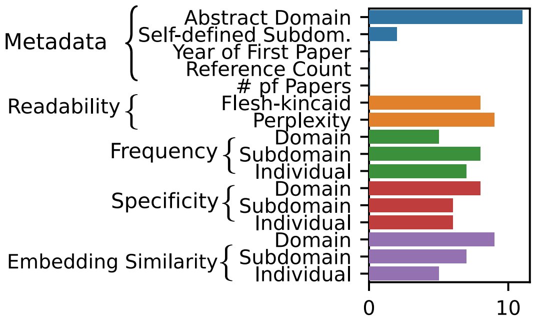


Figure 4: Frequency of non-zero coefficients in individual Lasso models across researchers. The Lasso penalty minimizes less critical coefficients to zero. Features with higher frequencies of non-zero values in individual models are consistently identified as important.

parison of precision, recall and F1 scores across the highest performing predictive models. Both oracles outperform all other methods, pointing to the benefit of using ratings from similar annotators for predicting term familiarity. The nearest-neighbor oracle achieves roughly the same performance as the majority oracle while using one tenth the data (i.e., one annotator rather than 10), suggesting that collaborative-filtering approaches could be effective for personalized familiarity prediction without collecting data from many annotators.

While other models slightly outperform the majority baseline, we generally see that the task of personalized jargon prediction is difficult. Among the models, the individual Lasso model shows superior performance compared to the mixed Lasso. Despite being provided less data, GPT-4 approaches slightly outperform the individual Lasso model.

Notably, the Lasso models achieve substantially higher precision (Table 2) than either GPT-4 or the majority baseline. Explaining terms a reader already knows (i.e., a false positive) distracted readers in our formative study (§3.1), making Lasso models that minimize false positives a promising alternative to prompt-based methods.

RQ2. What features and granularity level influence performance? Figure 4 reveals the non-zero coefficients of the individual Lasso models. The domain of the target abstract is consistently identified as a significant feature by all models. Word specificity and embedding similarity at the domain level also underscore the relevance of a researcher's domain in their familiarity with jargon, aligning with previous population-level familiarity research (Li et al., 2020). However, the importance of individual-level features such as frequency, specificity, and embedding similarity also emerge, highlighting the dual influence of both domain-specific and individual-specific factors.

Model	F1	Recall	Precision
Metadata			
Self-defined Subdom.	65.2$_{\pm 1.5}$	95.3$_{\pm 1.0}$	49.6$_{\pm 1.7}$
# of Papers	60.0$_{\pm 1.7}$	77.5$_{\pm 1.9}$	48.9$_{\pm 1.9}$
Reference Count	60.5$_{\pm 1.6}$	79.5$_{\pm 1.8}$	48.8$_{\pm 1.7}$
Year of First Paper	63.5$_{\pm 1.5}$	95.9$_{\pm 0.9}$	47.5$_{\pm 1.6}$
Abstract Domain	62.9$_{\pm 1.5}$	99.7$_{\pm 0.3}$	46.0$_{\pm 1.6}$
All Metadata	64.0$_{\pm 1.5}$	94.4$_{\pm 1.0}$	48.4$_{\pm 1.6}$
Granularity			
Domain	63.0$_{\pm 1.5}$	99.5$_{\pm 0.3}$	46.1$_{\pm 1.6}$
Subdomain	63.4$_{\pm 1.5}$	99.3$_{\pm 0.4}$	46.6$_{\pm 1.6}$
Individual	64.2$_{\pm 1.5}$	98.7$_{\pm 0.5}$	47.6$_{\pm 1.6}$
Example Number			
n=1	63.6$_{\pm 1.5}$	99.1$_{\pm 0.4}$	46.9$_{\pm 1.6}$
n=5	64.2$_{\pm 1.5}$	98.7$_{\pm 0.5}$	47.6$_{\pm 1.6}$
n=10	64.0$_{\pm 1.5}$	98.7$_{\pm 0.5}$	47.4$_{\pm 1.6}$

Table 3: Mean GPT-4 model performance (±std) on term familiarity prediction in the test set. Context-enhanced learning with individual level data is used for granularity and example number. Underlined models are reported in Table 2.

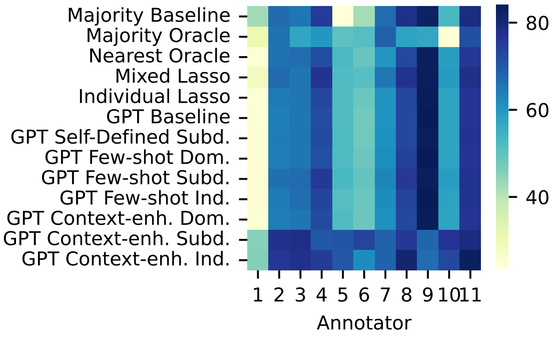


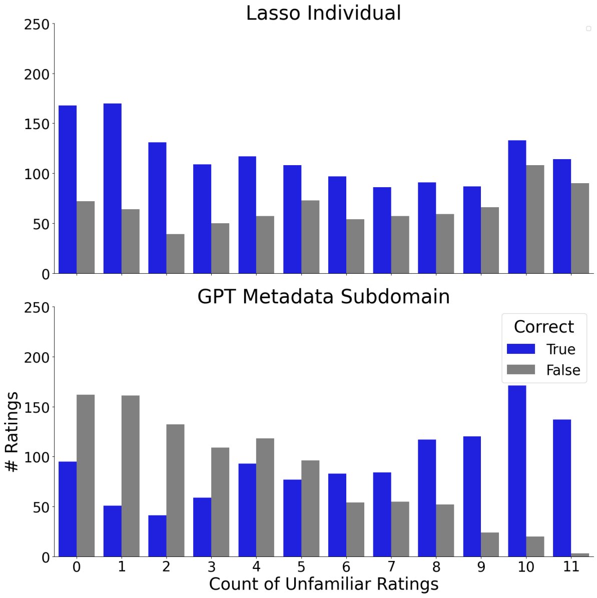
Figure 5: F1 score of models in familiarity prediction across annotators.

Table 3 details GPT-4 performance when including different metadata and publication granularities. Prompting with researcher self-defined subdomain seems to be the most effective strategy, suggesting that subdomain information (e.g., that a researcher is in NLP) is more useful than a researcher's broad field (e.g., CS).

RQ3. How do models perform across annotators? Model performance varies substantially for different annotators. Table 4 reports standard deviation of model performance calculated across annotators among top performing models and Figure 5 plots the F1 scores across annotators. We see ranges above 15 F1 points, well beyond the

Model	F1	Recall	Precision
Oracle			
Majority	±9.4	±13.2	±18.2
Nearest-neighbor	±10.4	±9.4	±16.8
Lasso			
Individual	±18.7	±25.8	±10.4
GPT			
Metadata - all	±15.6	±4.1	±17.5
Metadata - subdomain	±15.8	±4.1	±17.5
Context-enhanced	±16.4	±1.8	±16.8

Table 4: Standard deviation of the highest performing models' scores across annotators.


Lasso Individual

GPT Metadata Subdomain

Figure 6: Error counts for two best performing model variants; entities are binned by their total count of unfamiliarity ratings summed over annotators. Generally, the GPT-based method over-predicts unfamiliarity.

differences in performance we see across models. The high levels of variation measured across annotators indicate that models work much better for certain annotators over others. Looking at Figure 5, we also see that the best performing model is different for different annotators. For example, for Annotator 8, the *context-enhanced GPT-4* at the individual level performed best, while *individual lasso* performed better for Annotator 9.

RQ4. What errors do the models make? Figure 6 plots the rate of correct predictions for the two best performing model variants: GPT with subdomain metadata and the individual Lasso regression models. We see that the GPT method over-predicts unfamiliarity: incorrect predictions are usually for terms that most annotators were familiar with. Looking at these entities, most have common meanings that a reader could guess given context (e.g., "coffee production" or "food supply chains"). In contrast, the Lasso model generally performs better for highly familiar terms, but suffers for terms that are highly unfamiliar. Supporting this difference, we generally see that the individual lasso models perform better for annotators who were familiar with more entities, while the prompt-based methods performed better for annotators who rated more entities as unfamiliar.

Model	Def.	Bg.	Ex.
Majority Baseline	$44.4_{\pm 1.8}$	$37.2_{\pm 1.7}$	$31.5_{\pm 1.8}$
Oracle			
Majority	$50.7_{\pm 2.8}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Nearest-neighbor	$54.3_{\pm 2.4}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Lasso			
Mixed	$13.7_{\pm 3.6}$	$15.0_{\pm 5.1}$	$0.0_{\pm 0.0}$
Individual	$56.4_{\pm 2.8}$	$48.1_{\pm 3.3}$	$28.7_{\pm 3.9}$
GPT			
Baseline	$47.7_{\pm 1.8}$	$40.0_{\pm 1.9}$	$31.9_{\pm 1.8}$
Self-defined Subfield	$48.4_{\pm 1.9}$	$39.3_{\pm 1.9}$	$32.4_{\pm 1.8}$
Context-enhanced	$47.5_{\pm 1.8}$	$38.6_{\pm 1.9}$	$32.5_{\pm 1.8}$

Table 5: Mean F1 score (±std) on **additional information needs** (definition, background, and example) prediction in the test set. Recall and precision are in App. Table 8, 9, and 10.

RQ5. How does individual information contribute to additional information prediction? We investigate models that performed well on the familiarity prediction task to predict users' additional information needs. Results in Table 5 demonstrate that incorporating individual features into Lasso or prompt-based models do not substantially improve performance on this challenging prediction task. Predicting granular user information needs remains difficult without more tailored modeling and data to detect individual variations.

7 Discussion

This paper introduces the novel task of personalized scholarly jargon detection. We collect a dataset of over 10K term familiarity and information need annotations from 11 CS researchers. Our dataset reveals significant variation in term familiarity based on annotators' background (§4.4). When predicting familiarity, subdomain information and sampled paper abstracts written by the author can improve prediction. This is particularly evident in prompt-based methodologies, where the use of paper abstracts was more beneficial than term familiarity labels from that researcher. This might be because researcher abstracts/metadata provide more generalizable information about the annotator than individual-specific term familiarity labels.

The high levels of variation, of both familiarity labels (§4.4) and model performance (§6) across annotators, indicate that certain models work much better for certain annotators than others. Personalized scientific jargon identification is a difficult task, with no clear best modeling strategy, and performance is dependent on the individual researcher being modeled. This opens up new research questions and avenues for future work. For example, we see that collaborative oracles saw noticeable performance boosts, suggesting that taking ratings from similar researchers might improve over general-purpose jargon identification methods. Subdomain information was also helpful for prediction, suggesting that modeling familiarity at the subdomain level can bring benefits beyond current general jargon identification techniques. Furthermore, for researchers with a small publication history (and therefore a small amount of unlabeled individual-level data), subdomain data can provide valuable information for predicting jargon familiarity.

Our methods provide an exciting first step to support researchers in reading and communicating outside their domain (Wudarczyk et al., 2021). Combined with text simplification techniques (Srikanth and Li, 2020), models that can predict an individual researcher's familiarity could assist researchers by rewriting an abstract tuned to particular research audiences, a capability that our formative study findings highlighted as a need for researchers (§3).

8 Conclusion

This paper introduces the novel task of scientific jargon detection for the individual researcher. We collect and release a dataset of over 10K term familiarity annotations from computer science researchers and investigate supervised and prompt-based methods to predict term familiarity. We find that leveraging a researcher's publication history, self-reported subdomain, and general domain information can improve term familiarity prediction. Our results provide insight on integrating an individual's knowledge into scientific jargon detection.

Limitations

We focus on CS researchers for selecting annotators and abstracts in other fields. This might limit our ability to generalize to other domains or researchers. Publication venues (e.g., journals or conferences) and norms vary widely across fields, which might complicate how research publications can be used to model term familiarity. Because of the cost of collecting annotations, our dataset is also relatively small, which might further limit its generalizability. 11 CS researchers is not representative of all of CS nor does it adequately cover each subfield (e.g., only one annotator self-identified as a CS theory researcher). Our goal with the dataset and analysis is to show the potential of modeling individual term familiairty and information needs. We are excited about future work expanding this goal into new domains. Therefore, we published the procedures and questionnaire used in our data collection to encourage future research on personalized jargon detection for scientists in additional domains. Details can be found at: https://github.com/talaugust/PersonalizedJargon.

Ethics Statement

Some of the methods in the paper include personal data (e.g., publication record, labeled terms), which might pose a privacy risk for some researchers. Systems identifying term familiarity and information needs must keep any personal data stored locally and allow researchers to remove or view their data at any time. Focusing on CS researchers as a first step for predicting term familiarity might also allow CS researchers to more effectively read outside their discipline, but not researchers in other domains reading within CS. While encouraging interdisciplinary reading can improve two-way communication, it is also important to consider the voices of researchers in other domains.

Acknowledgements

We thank our participants and annotators as well as Bailey Kuehl for her valuable suggestions on annotation. We also thank our anonymous reviewers, the Semantic Scholar research team, as well as Raymond Fok, who read or provided feedback and assistance. This work was performed during an internship at AI2 and was supported in part by US National Library of Medicine [grant number R21LM013934].

References

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*, 30:1 – 38.

Adrian P. A. Barnett and Zoe Doubleday. 2020. The growth of acronyms in the scientific literature. *eLife*, 9.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Bernard C. K. Choi and Anita W. P. Pak. 2007. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. promotors, barriers, and strategies of enhancement. *Clinical and investigative medicine. Medecine clinique et experimentale*, 30 6:E224–32.

Kristy L Daniel, Myra McConnell, Anita Schuchardt, and Melanie E. Peffer. 2022. Challenges facing interdisciplinary researchers: Findings from a professional development workshop. *PLoS ONE*, 17.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *ArXiv*, abs/2310.07581.

Dee Gardner and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3):305–327.

David Goldberg, David A. Nichols, Brian M. Oki, and Douglas B. Terry. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35:61–70.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. *ArXiv*, abs/2205.02564.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor A. Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *ArXiv*, abs/2211.03818.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *ACM Conference on Recommender Systems*.

Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2020. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S. Weld, Doug Downey, and Jonathan Bragg. 2022. From who you know to what you read: Augmenting scientific recommendations with implicit social networks. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In *AAAI Conference on Artificial Intelligence*.

Hsin-Ni Lin, Shu-Kai Hsieh, and Shiao-Hui Chan. 2012. Measuring individual differences in word recognition: The role of individual lexical behaviors. In *ROCLING/IJCLCLP*.

Hongjun Liu, Sheng Dai, and De-en Jiang. 2014. Solubility of gases in a common ionic liquid from molecular dynamics based free energy calculations. *The Journal of Physical Chemistry B*, 118(10):2719–2725.

Li Lucy, Jesse Dodge, David Bamman, and Katherine A. Keith. 2022. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Annual Meeting of the Association for Computational Linguistics*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Alejandro Martínez and Stefano Mammola. 2021. Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society B*, 288.

Sonia K. Murthy, Daniel King, Tom Hope, Daniel S. Weld, and Doug Downey. 2021. Towards personalized descriptions of scientific concepts.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pontus Plaven-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *eLife*, 6.

Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLoS ONE*, 12.

Isabelle De Ridder. 2002. Visible or invisible links? In *CHI Extended Abstracts*.

Angelo Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. 2018. The computer science ontology: A large-scale taxonomy of research areas. In *International Workshop on the Semantic Web*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *ArXiv*, abs/2211.13308.

Neha Srikanth and Junyi Jessy Li. 2020. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings*.

Myra H. Strober. 2006. Habits of the mind: Challenges for multidisciplinary engagement. *Social Epistemology*, 20:315 – 331.

Kumiko Tanaka-Ishii and Hiroshi Terada. 2011. Word familiarity and frequency. *ArXiv*, abs/1806.03431.

Daril A. Vilhena, Jacob Gates Foster, Martin Rosvall, Jevin D. West, James A. Evans, and Carl T. Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, 1:221–238.

Olga A. Wudarczyk, Murat Kirtay, Anna K. Kuhlen, Rasha Abdel Rahman, John-Dylan Haynes, Verena V. Hafner, and Doris Pischedda. 2021. Bringing together robotics, neuroscience, and psychology: Lessons learned from an interdisciplinary project. *Frontiers in Human Neuroscience*, 15.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 377–386.

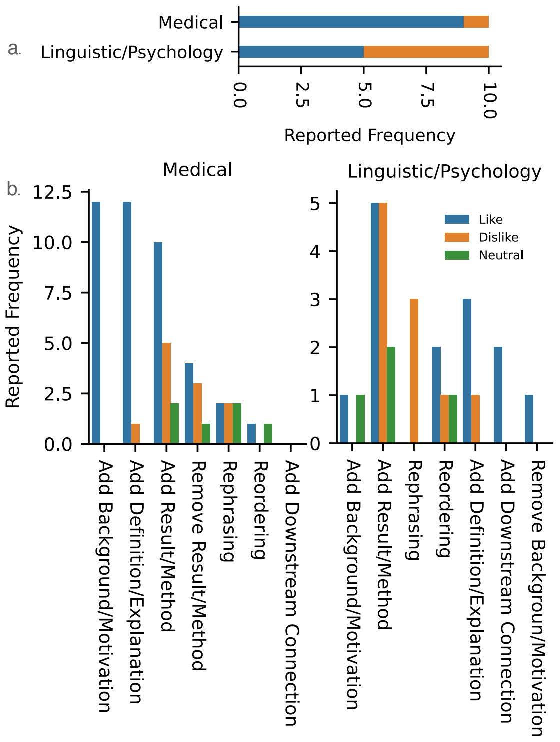


Figure 7: Formative study results: domain-specific attitudes towards (a) personalized abstracts and (b) transformations of personalized abstracts. Medical abstracts are distant from the annotators' background, whereas linguistic or psychology abstracts are closer. Legend is shared for both plots.

Appendix

A Initial Task Definition

For our initial study, we focus on reading full scientific abstracts because we wanted to understand if term familiarity was an important part of our broader envisioned setting of researchers reading abstracts outside of their domain. Further, our goal in providing a naive personalized abstract was to probe what transformations are feasible with current models that researchers respond positively to.

One abstract was from the medical domain, a domain not within any participant's dominant expertise, and one abstract was from a domain related to but distinct from a participant's specialization (i.e., psychology or linguistics). Personalized abstracts were generated using *text-davinci-003* model from OpenAI (OpenAI, 2023; Brown et al., 2020).

B GPT Prompts

All prompts used for GPT-4 experiments are shown in Table 7.

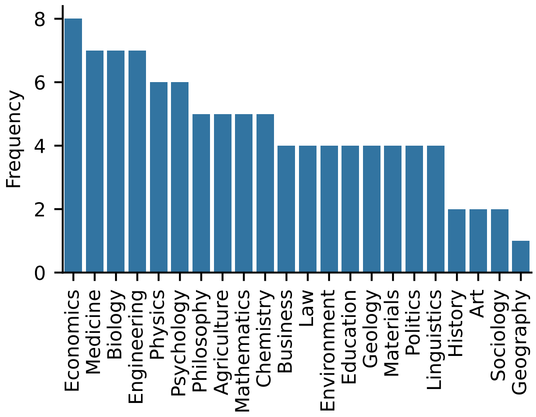


Figure 8: Number of papers categorized by domain in the annotation dataset.

Modification	**Original abstract sentence**	**Personalized abstract sentence**	**Human annotation**
Add Background/Motivation	Undiagnosed chronic kidney disease (CKD) is a common and usually asymptomatic disorder that causes a high burden of morbidity and early mortality worldwide.	Early detection of CKD is crucial for avoiding renal replacement therapy, **with an estimated 4.6% of total mortality worldwide**	*Like. I didn't realize how big the problem was, now I do.*
Add Definition/Explanation	Distributional semantics provides multidimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown by a large body of research in computational linguistics; yet, its impact in theoretical linguistics has so far been limited.	Distributional semantics is based on the Distributional Hypothesis, which states that similarity in meaning results in similarity of linguistic distribution.	*Like. Good to introduce what distributional semantics is.*
Add Result/Method	We developed a deep learning model for CKD screening from routinely acquired ECGs.	We also used **Local Interpretable Model-agnostic Explanations (LIME)** to identify which ECG segments were particularly used in the identification of CKD, which focused mostly on QRS complexes and PR intervals.	*Marking as neutral because this isn't in the abstract so I'm unsure if this is right. But if it is, I really like that it's highlighting more of the CS methodological stuff.*
Rephrasing	Undiagnosed chronic kidney disease (CKD) is a common and usually asymptomatic disorder that causes a high burden of morbidity and early mortality worldwide.	Chronic kidney disease (CKD) is a **major global health burden**.	*Like. simpler language.*
Reordering	Distributional semantics provides multidimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown by a large body of research in computational linguistics; yet, its impact in theoretical linguistics has so far been limited.	This survey provides an overview of the literature on distributional semantics, with a focus on methods and results that are of relevance for theoretical linguistics. Distributional semantics is based on the Distributional Hypothesis, which states that similarity in meaning results in similarity of linguistic distribution.	*Like. Not new but re-ordered; I like having this sentence at the start, as the first sentence in the original abstract is long and complicated.*
Add Downstream Connection	-	The paper also discusses how distributed representations can be used to generate dictionary definitions of words, and how they can be augmented with simple mathematical or logical expressions.	*Like. These additional details help make the study more concrete, and I understand them because this is my area. Also this connects to work I especially know because I've done some myself (generating dictionary definitions).*

Table 6: Example sentence pair for the original abstract and personalized abstract. Human annotation results for their attitudes and reasons towards the modifications. The personalized abstract was generated using the abstract personalization prompt in Table 7.

Task	Prompt	Model	Max length	Temp.
Abstract personalization	You are tasked with the role as a scientific writer to generate an personalized abstract for an individual reader. To do this effectively, consider including relevant background information or motivations related to the subject matter, provide necessary definitions or explanations to elucidate complex concepts, and incorporate significant methodological or result-oriented details. However, ensure that any additional information included is directly relevant and can be traced back to the provided content. The paper needed to be personalized is:{ }; The reader's publications are: { }; The reader's references are:{ }; The personalized abstract is: .	text-davinci-003	500	0
Top-10 significant terms	Please review the following scientific paper abstract. Your task is to identify all scientific-related word/phrases within the text and then rank these word/phrases in descending order based on their significance within the abstract itself. Retain the first 10 word/phrases:.	text-davinci-003	100	0
Familiarity classification	Your job is to estimate how much the reader knows about an entity. You will be provided with the entity, the abstract where the entity came from, and related data about either the reader or the abstract. Entity: { } Abstract:{ } Related Data:{ } Here's how to gauge the reader's familiarity: - 0: The reader knows this subject well and can describe it to others. - 1: The reader has either encountered this subject before but knows little about it, or has never come across it at all. Based on the information provied, determine the familiarity score, either 0 or 1:	gpt-4	100	0
Definition needs classification	Your job is to estimate whether the reader might need an additional definition to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity came from, and related data about either the reader or the abstract. Definition of definition/explanation: provides key information on the term independent of any context (e.g., a specific scientific abstract). A definition answers the question, Ẅhat is/are [term]?̈. Entity: { } Abstract:{ } Rel:{ }. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0
Background needs classification	Your job is to estimate whether the reader might need additional background to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity come from, and related data about either the reader or the abstract. Definition of background/motivation: introduces information that is important for understanding the term in the context of the abstract. Background can provide information about how the term relates to overall problem, significance, and motivation of the abstract. Entity: { } Abstract:{ } Rel:{ }. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0
Example needs classification	Your job is to estimate whether the reader might need additional example to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity come from, and related data about either the reader or the abstract. Definition of example: offers specific instances that help illustrate the practical application or usage of the term within the abstract. Entity: { } Abstract:{ } Rel:{ }. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0

Table 7: GPT prompts and configurations. We prompted models in September of 2023.

Model	**F1**	**Recall**	**Precision**
Majority Baseline	44.4$_{\pm 1.8}$	100.0$_{\pm 0.0}$	28.6$_{\pm 1.5}$
Oracle			
Majority	50.7$_{\pm 2.8}$	42.7$_{\pm 2.9}$	62.3$_{\pm 3.4}$
Nearest-neighbor	54.3$_{\pm 2.4}$	60.4$_{\pm 2.9}$	49.3$_{\pm 2.7}$
Lasso			
Mixed	13.7$_{\pm 3.6}$	24.0$_{\pm 6.2}$	9.6$_{\pm 2.7}$
Individual	56.4$_{\pm 2.8}$	45.5$_{\pm 3.0}$	74.1$_{\pm 3.4}$
GPT			
Baseline	47.7$_{\pm 1.8}$	97.3$_{\pm 1.0}$	31.6$_{\pm 1.6}$
Self-defined Subfield	48.4$_{\pm 1.9}$	92.9$_{\pm 1.6}$	32.7$_{\pm 1.6}$
Context-enhanced	47.5$_{\pm 1.8}$	97.0$_{\pm 1.1}$	31.5$_{\pm 1.6}$

Table 8: Mean model performance (±std) on **additional definition** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category. Standard deviation is calculated by bootstrapping.

Model	**F1**	**Recall**	**Precision**
Majority Baseline	37.2$_{\pm 1.7}$	100.0$_{\pm 0.0}$	22.8$_{\pm 1.3}$
Oracle			
Majority	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
Nearest-neighbor	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
Lasso			
Mixed	15.0$_{\pm 5.1}$	10.5$_{\pm 3.8}$	26.3$_{\pm 8.7}$
Individual	48.1$_{\pm 3.3}$	36.9$_{\pm 3.2}$	69.3$_{\pm 4.0}$
GPT			
Baseline	40.0$_{\pm 1.9}$	95.5$_{\pm 1.3}$	25.3$_{\pm 1.5}$
Self-defined Subfield	39.3$_{\pm 1.9}$	93.1$_{\pm 1.7}$	24.9$_{\pm 1.5}$
Context-enhanced	38.6$_{\pm 1.9}$	96.9$_{\pm 1.2}$	24.1$_{\pm 1.4}$

Table 9: Mean model performance (±std) on **additional background** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category.

Model	**F1**	**Recall**	**Precision**
Majority Baseline	31.5$_{\pm 1.8}$	100.0$_{\pm 0.0}$	18.7$_{\pm 1.2}$
Oracle			
Majority	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
Nearest-neighbor	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
Lasso			
Mixed	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$
Individual	28.7$_{\pm 3.9}$	18.0$_{\pm 2.8}$	71.7$_{\pm 6.6}$
GPT			
Baseline	31.9$_{\pm 1.8}$	97.4$_{\pm 1.2}$	19.1$_{\pm 1.2}$
Self-defined Subfield	32.4$_{\pm 1.8}$	91.3$_{\pm 2.1}$	19.7$_{\pm 1.3}$
Context-enhanced	32.5$_{\pm 1.8}$	97.4$_{\pm 1.1}$	19.5$_{\pm 1.3}$

Table 10: Mean model performance (±std) on **additional example** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category.

Annotator		Familiarity		Definition		Background		Example	
ID	Sub-field	Neighbor	Sub-field	Neighbor	Subfield	Neighbor	Subfield	Neighbor	Subfield
1	CV	10	CV	5	Security	8	NLP	4	NLP
2	Networking	3	NLP	9	Networks	3	NLP	11	CV
3	NLP	11	CV	4	NLP	11	CV	2	Networking
4	NLP	11	CV	3	NLP	11	CV	1	CV
5	Security	2	Networking	2	Networking	11	CV	2	Networking
6	Theory	3	NLP	9	Networks	3	NLP	9	Networks
7	NN	8	NLP	9	Networks	4	NLP	10	CV
8	NLP	11	CV	3	NLP	11	CV	11	CV
9	Networks	7	NN	2	Networking	4	NLP	6	Theory
10	CV	3	NLP	3	NLP	8	NLP	5	Security
11	CV	3	NLP	4	NLP	8	NLP	2	Networking

Table 11: Nearest-neighbor for each annotator.

Entity sentence	**Domain**	**Ann. subdomain**	**Fam.**	**Information**	**Total Fam.**
We study auctions for selling a limited supply of a single **commodity** in the case where the supply is known in advance and the case it is unknown and must be instead allocated in an online fashion.	Economics	CS Theory	Familiar	Example	9
Inference with the graphical model for de novo peptide sequencing estimates **posterior probabilities** for amino acids rather than scores for single symbols in the sequence.	Engineering	Computer Vision	Familiar	Example	7
… to communicate foreign policy goals and decisions, construct a **strategic narrative** of Indian foreign policy and counter narratives inimical to Indian interests.	Political Science	NLP	Familiar	None	6
Several parameters obtained from the experimental results were compared and analyzed, including the **load-bearing capacity**, stiffness, ductility, energy dissipation, and failure characteristics of the specimens.	Engineering, Materials Science	CS Theory	Unfamiliar	Definition	5
Agents with identical **linear time-invariant dynamics** are considered.	Mathematics	Neural Networks	Familiar	None	2
Self-directed learning is a necessary skill for students and workers to remain lifelong learners.	Education	NLP	Unfamiliar	Definition, Motivation	9

Table 12: Sample of entities with sentence context, their familiarity ratings, and information needs. Entities are bolded within the sentence. Total familiarity count indicates how many annotators rated an entity as familiar.