Using Open Data to Automatically Generate Localized Analogies

Sofia Eleni Spatharioti
Microsoft Research
New York, New York, USA
s.spatharioti@gmail.com

Daniel G. Goldstein
Microsoft Research
New York, New York, USA
dgg@microsoft.com

Jake M. Hofman
Microsoft Research
New York, New York, USA
jmh@microsoft.com

ABSTRACT

Numerical analogies (or “perspectives”) that translate unfamiliar measurements into comparisons with familiar reference objects (e.g., “275,000 square miles is roughly as large as Texas”) have been shown to aid readers’ recall, estimation, and error detection for numbers. However, because familiar reference objects are culture-specific, analogies do not always generalize across audiences. Crowdsourcing perspectives has proven effective but is limited by scalability issues and a lack of crowdworking markets in many regions. In this research, we develop an automated technique for generating localized perspectives. We utilize several open data sources for relevance signals and develop a surprisingly simple model capable of localizing analogies to new audiences without any retraining from human judges. We validate the model by testing it in both a new domain and with a different linguistic audience residing in another country. We release the compiled dataset of 400,000 reference objects to the research community.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

perspectives,numeracy,measurement,crowdsourcing

ACM Reference Format:
Sofia Eleni Spatharioti, Daniel G. Goldstein, and Jake M. Hofman. 2024. Using Open Data to Automatically Generate Localized Analogies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3613904.3642638

1 INTRODUCTION

News stories are rich with statistics and numerical measurements, many of which are difficult to comprehend. Consider the following examples: The estimated cost of damages for Hurricane Isaac, which struck North America in 2012, was $3 billion [3]. The year 2021 was a devastating year for wildfires in Europe, with 1.1 million hectares of area consumed [35]; the area burned in Greece alone was 131,254 hectares [7]. Over two million Americans live without regular access to piped water [26].

Figures like $3 billion dollars, 1.1 million hectares, 131,254 hectares, and two million people do not look out of place in the news, but both academic research and popular accounts suggest that readers often have difficulty comprehending them or detecting if they are wildly incorrect [2, 4, 17, 30]. A solid understanding of magnitudes and measurements is important for voters, consumers, and policy makers [5, 10, 21].

An idea that was proposed and validated in the HCI literature is to improve the communication of unfamiliar measurements by expressing them through analogies, also known as *perspectives* [2, 16, 18, 19, 34]. Perspectives employ *reference objects* that are broadly familiar to an audience and easy to understand and visualize. In the above examples, it might be difficult for someone in the U.S. to imagine 1.2 million hectares, but would have an easier time doing so if they were told that 1.2 million hectares is about half as large as New Jersey. The affected area in Greece in 2021 can be similarly expressed as about equal to the size of the city of Los Angeles. Finally, we can visualize the two million Americans without piped water by evoking the thought of the entire borough of Manhattan lacking piped water.

While expressing unfamiliar measurements using helpful analogies has been shown to improve reader comprehension [2], identifying which are the best reference objects for a particular measurement remains a challenge. Prior approaches have relied on crowdsourcing reference objects and building domain-specific models for each national and linguistic audience [16, 19, 34]. While this can be effective for surfacing high quality reference objects, it can be difficult to scale such techniques to cover a wide range of measurements. Moreover, the characteristics of the crowd involved in the process may influence the types of reference objects that are generated. For example, a U.S.-based crowd may consider Manhattan as a good analogy for two million people, but this analogy may not be as useful for audiences outside of the United States. A better analogy for French or European audiences might be Paris. Consequently, adapting analogies to different audiences and cultures may require a separate crowdsourcing process for each audience.

To address these challenges, we develop an automated approach to generating localized numerical analogies based entirely on open data sources. We do so by first constructing a large database of reference objects extracted from the Wikidata knowledge graph, which we make available for public use. For each of these reference objects, we collect relevance signals from several additional sources, including search activity, word frequencies, text embeddings, and Wikipedia traffic. We then evaluate a series of models using these signals to predict a small set of relevance judgements from U.S.-based crowd workers. Interestingly—and not at all obvious *a priori*—in an ablation study we find that Wikipedia traffic alone is a reliable predictor of relevance, and that adding other signals does not significantly boost model performance. This allows us

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642638

to use a single, simple model to localize analogies to new domains and new audiences without any retraining from human judges. We evaluate the approach on both a new domain (with U.S. judges) and with a new audience (French judges), finding that it performs well on these challenging, out-of-distribution tests.

In the remainder of the paper, we first review the literature related to numerical analogies and using open data for relevance signals. We then present details of the dataset and model we constructed. We conclude with a discussion of key takewaways from our work, along with limitations and directions for future research.

2 BACKGROUND

2.1 Perspectives for Measurements

An effective means to assist individuals in understanding unfamiliar measurements is through graphical representations [31] or Virtual Reality [22]. In our study, we translate unfamiliar measurements using textual analogies. Despite their lack of visual appeal, these analogies can be effectively communicated through written or spoken language.

Formulating easily comprehensible text analogies is a difficult task, necessitating an understanding of the qualities that make an object widely recognized as familiar and useful. Two primary strategies have been pursued thus far: one employs crowdsourcing to establish non-personalized reference objects, while the other generates individualized analogies via predefined rules.

The first strategy is exemplified by the work of Barrio et al. [2], who devised a crowdsourcing method based on templates to elicit reference objects. The perspectives were used in series of studies demonstrating their ability to enhance the numerical comprehension of news stories. In a similar vein, Riederer et al. [34] proposed a statistical model for generating perspectives for population and area measurements. This model takes into account crowd participant estimates, the effect of multipliers (e.g., "half as large", "five times as large"), and object familiarity. More recently, perspectives have been employed to improve understanding of the carbon footprint associated with various travel options [29]. Crowdsourcing indeed holds potential for producing high-quality perspectives due to its ability to tap into a crowd's collective sense of what is familiar and useful. Nonetheless, scalability issues arise with this approach, such as accommodating a broad spectrum of values and measurements, or identifying suitable objects for diverse audiences not well-represented by crowd workers. Closer to our work, Hullman et al. [16] consider object and measure familiarity in generating re-expressions, with familiarity determined using WordNet, ImageNet and crowdsourcing techniques. The crowdsourced generation of analogies using templates has also been explored in the field of explainable artificial intelligence [14], where the quality of analogical properties has been conceptualized as relating to structural correspondence, relational similarity, transferability and helpfulness. Work in this area has noted both the difficulty of generating consistently high quality analogies through crowdsourcing, as well the subjective quality of the analogies themselves. As different people prefer different analogies, there seems to be no one-size-fits-all solution.

The second approach is exemplified by the work of Kim et al. [19], in which distances, for example, are contextualized using landmarks and points of interest relative to the reader's location. In this approach, reference objects are sorted using a domain-specific model with a manually-specified ranking function.

Our goal is to strike a balance between the non-personalized approach, which caters to a single audience, and the individualized approach, which may not always be viable due to insufficient user-specific information. We aim to construct reference objects tailored to specific audiences or cultures. Instead of relying heavily on crowdsourcing for reference objects, we utilize it primarily to train a model that draws predominantly on open data sources such as Wikidata and Wikipedia. The resulting model can then exceed the capabilities of crowdsourcing in terms of both the number and variety of objects generated, and the diversity of audiences it can serve.

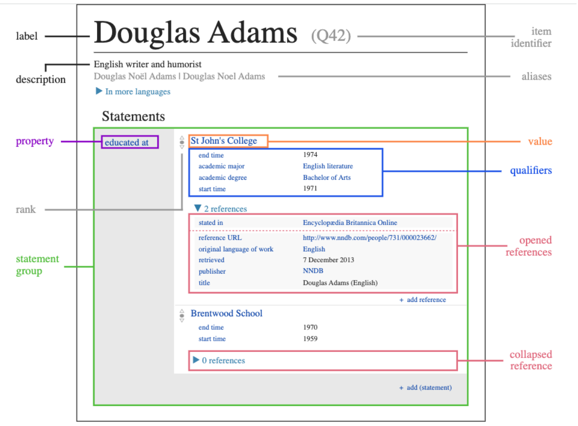
2.2 Analogies for Statistical Indicators

The use of analogies to improve understanding of statistical indicators in the form of probabilities and percentages has been heavily explored in the literature. Barilli et al. explored communicating probabilities related to risk using verbal analogies [1]. In particular, they studied different levels of risk and the presence of verbal analogies in the form of the ball and urn example. Participants' risk perceptions were reduced when using these analogies, due to a focus more on the re-expression (e.g., balls in an urn) than the original health related risk. These findings highlight the importance of carefully evaluating analogies in health settings to ensure they do not inadvertently backfire.

Galesic and Garcia-Retamero surveyed participants across different numeracy levels and countries of origin on their comprehension of risks [12]. Re-expressing consequences of health-related behaviors as increases or decreases in life expectancy led to higher recall levels, compared to using the probability of contracting a disease. Here, using analogies did not bias risk perceptions. Encoding and recall were both enhanced when information was easier to imagine.

A separate set of studies related to [12] examined the effect of various analogies in risk comprehension through a series of experiments [13]. Groups of people were tasked with rating analogies for similarity and familiarity. Different analogies were helpful for different audiences, and overall using analogies in difficult tasks improved the performance of high-numeracy participants in two countries examined. There was an interaction between numeracy skills and problem difficulty, pointing to potential negative effects of presenting analogies to people who already have a good understanding of the measurement.

Analogies have proven useful in helping people understand statistical concepts. Martin [25] provides a comprehensive overview of a range of analogy strategies deployed to help students understand concepts such as statistical plots, variance, standardized scores, probability, distributions, hypothesis testing, and more. Kim et al. [18] test the effectiveness of analogies that put statistical effect sizes in perspective, focusing on ones that can be understood by a wide variety of people (such as comparisons of heights of people at various ages).



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

▶ In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

▶ 0 references

+ add (statement)

Figure 1: Example of a Wikidata Entry. Image by Charlie Kritschmar (WMDE) - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=49616867

2.3 Relevance signals

In this work, we study four different proxies for relevance: search volume, Google N-grams, text embeddings and Wikipedia pageviews, and generate perspectives using Wikidata entities. Search volume information via Google Trends has been utilized in linking public interest on biodiversity conservation efforts to zoo programs and the airing of an animated TV program from 2011 to 2018 in Japan [11]. Information sourced from the Baidu search engine has also been explored for predicting popularity of table tennis players [24], while N-grams and embeddings have also been effective proxies for predicting the popularity of particular headlines [20]. We find that search volume, N-grams and embeddings are relatively ineffective for generating perspectives for numerical information, and identify specific drawbacks for each of these approaches.

Wikipedia traffic is often triggered by events circulating in the media or social discussions [36], making it a potent tool for approximating popularity and familiarity. Wikipedia pageviews have been extensively explored in a variety of settings [8, 28, 32, 37, 39, 40]. In addition, elements such as content diversity, cultural contextualization and heterogeneous motivations exhibited across different Wikipedia language editions [15, 23, 27] hint at this signal's potential for capturing audience-specific trends. Our work contributes to the literature on the power of Wikipedia pageviews. We find that this type of proxy can lead to better performance in generating helpful perspectives, and bears characteristics that make it particularly suitable for adapting perspective suggestions to different audiences.

3 BUILDING A RELEVANCE MODEL

The process of developing models for automatically generating localized perspectives consists of the following main steps: First, we construct a dataset of candidate reference objects that can be used to describe a variety of measurements. Next, we identify information that can be collected for these objects that may serve as proxies for helpfulness and familiarity. These proxies serve as the features of our model. Finally, we analyze the effectiveness and associated cost of the different proxies to narrow down the features for our final model. We consider only analogies that can be constructed using objects that are approximately the same size as the measurement, motivated by prior work on the effectiveness of multipliers of 1 [34] on comprehension and accuracy.

In what follows we describe everything in terms of English-based sources. In theory the same procedures can be conducted in other languages or for other cultures, but as we will show, the method we present alleviates the need for language- or culture-specific model development.

3.1 Reference Object Candidates

For identifying the initial dataset of candidate reference objects, we used Wikidata[1]. Wikidata is a free, open-sourced knowledge base, that is actively maintained and verified by Wikidata users. Wikidata offers a standardized structure and an open API that can be used to retrieve entities with a wide variety of measurements recorded. In addition, Wikidata is multilingual, which allows for retrieving language-specific information, such as localized names, descriptions, and links to external resources, such as language specific Wikipedia pages.

Using the Wikidata SPARQL API, we queried Wikidata for items that have some measurement attached. We initially focused on four core attributes; length, height, area and mass. For each core attribute, we constructed queries to fetch objects that may contain records for these attributes, in the form of different units.[2] For each object, we captured the following information:

- **Object Name**: The label (title) of the entity.
- **Wikidata ID**: The unique Wikidata identifier of the entity.
- **Attribute**: The attribute for which we have collected a measurement, e.g. height, length, mass etc..
- **Amount**: The numerical value of the recorded measurement.
- **Unit**: The reference unit for the recorded measurement.
- **Wikipedia URL**: The Wikipedia page identifier linked to the Wikidata entity.
- **Instance**: The class of which this Wikidata entity is recorded to be a particular example and member.
- **Countries**: A list of countries (if any) attached to the entity.

An example of a Wikidata entry can be found in Figure 1. Note that an individual Wikidata entry may have measurements for multiple attributes or even multiple measurements for a specific attribute. We treated cases of Wikidata entries with multiple attributes by creating individual records for each attribute. In cases where multiple measurements existed for a particular attribute, we picked the measurement with the highest rank or latest date, if a date was attached to the measurement. For the remaining cases with multiple measurements, we converted to the same unit and then picked the median measurement as the singular measurement. Finally, we rounded the measurements to two significant figures.

[1] https://www.wikidata.org/

[2] To respect query limitations while retrieving as many objects as possible, we used different approaches such as sub-queries in ranges for each attribute and each different unit, and time limits. We note that we did not perform an exhaustive collection of Wikidata entities at this stage, but subsequently extended the database, as we describe later in the paper.

In order to further narrow down a subset of items that could be representative of potential candidates for reference objects, we employed the following strategies. First, we consolidated cases where instances referred to the same type of object, for example multiple different names of instances describing ships. We identified and excluded objects that belonged to very obscure and specific instances, such as "hepogeum", "sheep breed" etc. Our initial dataset contained a high amount of entries related to humans (28%), such as the weight of Lisa Kudrow. We exclude these entries from further analysis, as they would not be of specific value for use in perspectives. Cleaning up and consolidating items resulted in a preliminary set of 39,532 Wikidata objects, belonging to 60 distinct Wikidata instance categories.

3.2 Proxies for Relevance

Having determined an initial set of candidate reference objects, we proceed to evaluate possible proxies for helpfulness and familiarity. We considered the following four main signals:

- **Search Volume (Bing)**: The estimated total number of available results when querying the object name via the Bing search engine API[3].
- **Ngram Counts (Ngram)**: Latest available Google Ngram Counts expressed as normalized frequencies.
- **Text Embeddings (BERT)**: Sentence Embeddings using SBERT Networks, using the paraphrase-mpnet-base-v2 model [33].
- **Wikipedia Pageviews (Wiki)**: The average number of monthly views of the English Wikipedia page of the item, and the number of years the Wikipedia page has been active.

A clear benefit of using **Search Volume** information as a proxy for helpfulness relates to its broad coverage and large scale. However, such information is also associated with two major drawbacks. We refer to the first drawback as *Entity Ambiguity*: For two independent entities with very similar or even identical names, discerning which of the information refers to which object can be very difficult. Consider Wikidata entities Q30[4] and Q47499411[5] as an example. The first refers to *United States of America*, the country, while the second refers to *America*, a 2016 sculpture depicting a fully functioning toilet made of solid gold by artist Maurizio Cattelan. Querying for the estimated total number of results for the second item may yield incorrect results, as some, or most pages may relate to the first entity. Mitigating entity ambiguity is especially challenging in "open-domain" tasks [9]. Search volume may also suffer from data quality issues, due to the nature of the metric (estimate).

A second category we examine is **Ngram Counts**, via Google Ngrams. Ngrams offer access to a familiar, well-defined corpus of books. For this category, we look at Ngram frequency, i.e. the percentage of occurrences of a given reference object name, out of all possible n-grams of the same size. For simplicity, we examine the most recent Ngram frequency for each object. We used the American English 2019 corpus. We note that Ngrams may also suffer from Entity Ambiguity issues, and also have a medium computation cost.

We further considered **Text Embeddings** as a potential proxy for helpfulness and familiarity, using the popular SentenceTransformers BERT Python framework. We used the paraphrase-mpnet-base-v2 pretrained model to encode each reference object into a list of 768 features. The high-dimensionality of text embeddings can offer a powerful solution that has been shown to have great performance in a variety of modeling tasks. However, we still have to deal with entity ambiguity issues, at a high computation cost.

A final category we examined is **Wikipedia Pageviews**. More specifically, we calculated the average number of pageviews the Wikipedia article attached to the Wikidata entity received each month, across its entire life span. In contrast to the previous categories, Wikipedia pageviews offer a concrete solution to the Entity Ambiguity challenge, as there is a direct, disambiguated link between reference object and proxy, guaranteeing the highest level of precision. Another benefit identified relates to the ability to collect Wikipedia pageviews in multiple languages, based on the number of available articles. While this is possible, to some extent, in all previous methods, there are a relatively small number of languages supported by other methods (i.e., Google Ngrams is available in 9 languages[6], Bing Search in 40 languages[7], and SBERT in ≈50 languages[8]). On the contrary, Wikipedia benefits from a global network of contributors, spanning articles in 322 currently active editions[9]. A comparison overview of all proxies can be found in Table 1.

3.3 Model Fitting and Ablation Study

With a database of reference objects and relevance signals, we proceeded to collect a small set of crowdsourced judgements to build and evaluate a series of relevance models. As collecting human assessments for the entire dataset of about 40,000 objects was infeasible, we aimed at creating a subset with some level of diversity and coverage in terms of objects. Specifically, we created a stratified sample based on **instance categories**. For each of the distinct instance categories in our set (60), we selected the top 5 items in terms of popularity, as measured in Wikipedia average monthly views. This strategy minimizes issues arising from the presence of dominant categories in our original dataset (e.g. countries, states and cities for area), while also filtering out encyclopedic entries that could not realistically be used as perspectives. This process resulted in 300 reference objects as training examples. Examples of reference objects can be found in Table 2.

To collect human assessments for these objects, we designed a task where we invited participants to rate reference objects for *helpfulness* and *familiarity.* For helpfulness, we showed participants a sentence comparing a measurement to a reference object, and asked them to rate how helpful this comparison is for understanding the measurement, on a 1 through 5 scale, with 1 being "Not at all", and 5 being "Extremely". For familiarity, we asked participants to rate their familiarity with the measurement of the reference object, again on a scale from 1 to 5, with 1 being "Not at all" and 5 being

[3] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api
[4] https://www.wikidata.org/wiki/Q30
[5] https://www.wikidata.org/wiki/Q47499411
[6] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html
[7] https://learn.microsoft.com/en-us/azure/cognitive-services/bing-news-search/language-support#supported-markets-for-news-search-endpoint
[8] https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models
[9] https://en.wikipedia.org/wiki/List_of_Wikipedias

Type	Source	Pros	Cons
Search Volume	Bing API	• Broad Coverage • Large Scale	• Entity Ambiguity • Data Quality Issues
Ngram Counts	Google Ngrams	• Familiar, well-defined corpus	• Entity Ambiguity • Medium computation cost
Text Embeddings	SBERT	• Broad Coverage • High-Dimensional	• Entity Ambiguity • High Computation Cost
Wikipedia Pageviews	Wikipedia	• Precise Entities • Cross-Cultural	• Medium-Low Computation Cost

Table 1: Comparison of features explored for modeling helpfulness and familiarity.

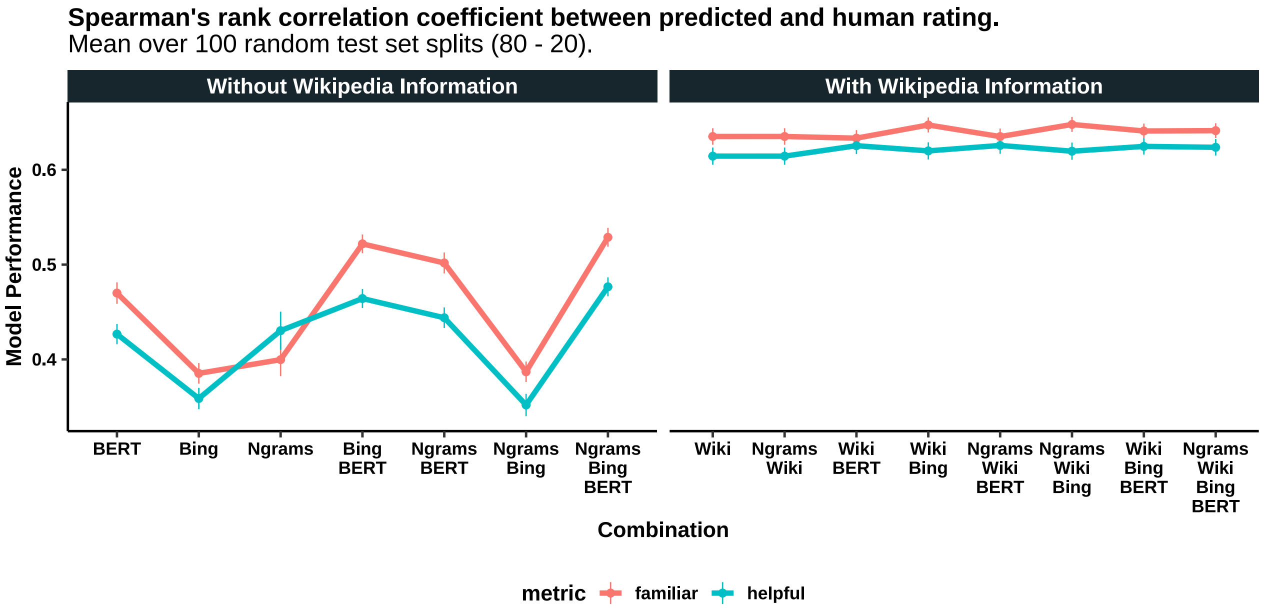


Figure 2: Results of Model Performance for all possible feature combinations. Error bars depict one standard error. We found that models containing Wikipedia information performed best overall, and that a simple model with just Wikipedia information had comparable performance to models with additional features.

Attribute	Reference Object	Instance Category	Measurement
Area	New York City Rajaji National Park	City National Park	470 square miles 110 square miles
Height	One World Trade Center Milan Cathedral	Skyscraper Church	1800 feet 360 feet
Length	Las Vegas Strip Ever Given	Street Ship	4.2 miles 0.25 miles
Mass	Voyager 1 Quarter	Space Probe Coin	1800 pounds 0.0097 pounds

Table 2: Examples of candidate reference objects, for different dimensions, as sourced from Wikidata.

“Extremely”. We recruited 70 U.S. based participants through Amazon Mechanical Turk, using the Masters qualification. Participants were paid $1.75 for completing the study. Participants were asked to rate 20 reference objects, randomly drawn from the pool of 300 objects, to receive payment. For each object, we calculated *helpfulness* and *familiarity* scores as the median of all human ratings for that object.

For the purposes of evaluating model features, we focused only on objects for which we were able to collect information for all four proxies and only considered reference objects for which we had received at least 3 ratings, resulting in a set of 245 objects. For generating rating predictions, we fit a cross-validated linear model with L1 regularization using the `glmnet` package in R. Search volume, Ngram counts, and Wikipedia pageviews were log-transformed to account for skew and then used as linear features in the corresponding models, whereas BERT embeddings were used directly as linear factors. We compared all possible combinations of candidates proxies (Bing, Ngram, BERT, Wiki) by conducting 100 random training-test splits and averaging the Spearman's rank correlation coefficient between predicted and human ratings.

A summary of results can be found in Figure 2. Models with Wikipedia pageviews outperform models without them, as can be seen by comparing the left panel of Figure 2 to the right panel. Surprisingly, we also observed that a simple model with *only* Wikipedia information achieved comparable performance to models with additional features for both the helpfulness and the familiarity metrics, as shown by the relatively flat trends in the right panel of Figure 2. From this we concluded that Ngram, Bing and BERT information can be safely ignored, greatly simplifying and lowering the transaction costs of the modeling process. Finally, we average the helpfulness and familiarity predicted ratings into one *model rating* and re-fit a model over all of the labelled data, which we used for subsequent evaluations. The predictive model used for the remainder of this paper is `glm: total_rating ~ log(wikipedia_page_monthly_views) + log(wikipedia_page_active_years)`.

4 MODEL EVALUATION

To assess the effectiveness of our trained model, we conduct two sets of evaluation studies to answer:

(1) How effectively does the trained model scale to **different measurement types**?
(2) How effectively does the trained model tailor examples to **global audiences**?

We use the same model throughout, trained on four measurement categories (area, mass, height, length) using Wikipedia pageviews, with no additional retraining.

4.1 Measurement Adaptability

To evaluate the model's capabilities in covering a broad range of measurements, we consider two cases. First, we consider new measurements, whose categories **are known to the model**, i.e. measurements for area, mass, height and length. Second, we consider a separate category of measurements for which **the model has not been trained on**: measurements for populations.

We proceed to create the following *evaluation subset*. For each of the 5 main measurement categories, we define 5 amount bins to sample objects from. This ensures our resulting subset covers a broad range of amounts across all measurements. For each bin, we picked the top object in the 100^{th}, 67^{th} and 33^{rd} model rating percentile respectively. to represent high, medium, and low quality analogies. This ensures our evaluation subset consists of objects of variable quality. This process yielded a total of $5 \times 5 \times 3 = 75$ analogies.

We published a HIT on Amazon Mechanical Turk to collect human ratings for these analogies, aiming for 20 ratings for each item, using the same interface described in the previous section. We recruited 100 U.S. based participants, using the Masters qualification as an additional quality measure. Participants were paid $2.5 for completing the study. Similar to the model rating, we defined the *human rating* of an object as the combination of the average helpfulness and familiarity scores. We plot the relationship between Model and Human ratings by attribute in Figure 3 and measure performance using Spearman Rank Correlation (ρ).

4.1.1 *Results.*

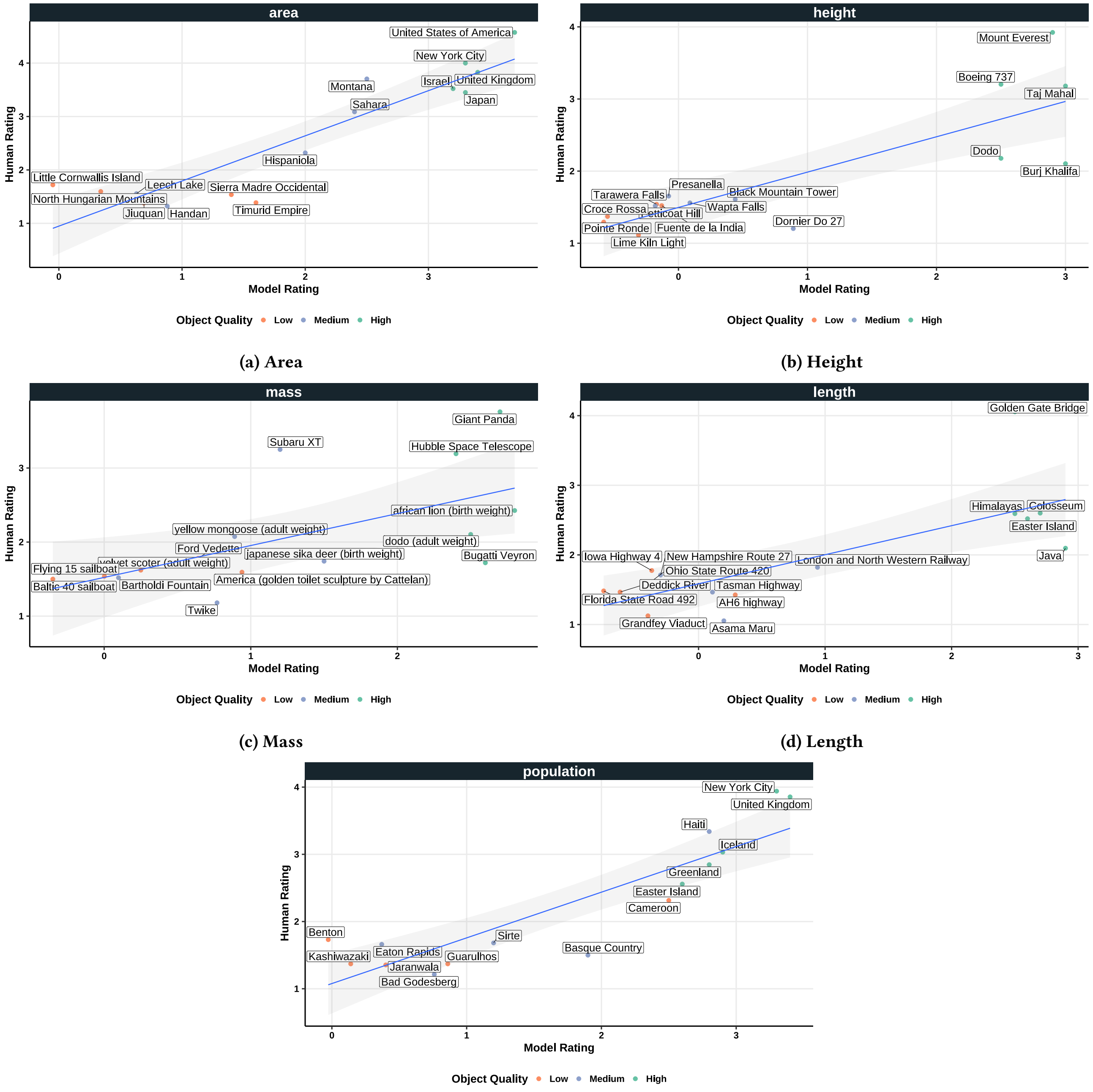
Despite the relatively small set of human judgements on which it was trained and the simple feature set used, our model achieved an overall $\rho = 0.77$. Looking at individual attribute categories, we observed that population, the category for which we effectively had no training samples, performed quite well at $\rho_{population} = 0.83$, highlighting the ability of the approach to generalize to new domains. Examples of good perspectives generated by our model include New York City and United Kingdom (Figure 3e). Our model is able to effectively filter out lower rated objects such as Kashiwazaki and Benton. We found the second highest performance when generating perspectives for area, with $\rho_{area} = 0.80$. The performance for height and mass was $\rho_{height} = 0.79$ and $\rho_{mass} = 0.74$ respectively. Finally, we saw the lowest performance in length, with $\rho_{length} = 0.64$. We do, however, note our model's ability to highlight the top rated items (Figure 3d).

4.2 Audience Adaptability

Modeling helpfulness and familiarity using Wikipedia information offers a potentially crucial advantage. Wikipedia is a cross-cultural, globally accepted knowledge base, with over 322 currently active editions, in multiple languages. This core characteristic allows access to an additional level of granularity for our model design, as we are able to collect pageviews for a Wikidata entity depending on the different available Wikipedia editions. Hence, we may be able to model familiarity and helpfulness of an object across different countries and cultures.

In this section, we explore the effectiveness of our model towards adapting to different audiences, by examining crowd preferences between the U.S. and France. More specifically, we designed a task in which participants from both countries were presented with a measurement and two reference objects, and asked to pick the one that would be most helpful to them in understanding the measurement. For each measurement, we picked the choices to reflect the top ranked reference object tailored for each audience, by generating ratings using our model based on the Wikipedia edition most closely related to that audience (i.e. English Wikipedia for U.S. and French Wikipedia for France).

To generate a reasonable list of measurements for this study, we again applied a stratified sampling strategy to ensure broad measurement coverage. We split each attribute in 5 bins, and for each attribute-bin pair we selected the measurement with the most objects in our Wikidata pool, resulting in $5 \times 5 = 25$ potential measurements (e.g. mass of 2100 pounds, height of 11000 feet etc.). To determine the top reference object per audience for each measurement, we first retrieved Wikipedia pageviews for all potential reference objects in our pool that could be used to describe this measurement. We then used our model to generate two rankings for each object, one for the French audience using the French Wikipedia pageviews, and one for the U.S. audience using English Wikipedia pageviews. We broke ties, determined as objects having overlap in their confidence intervals, by prioritizing objects with a direct relation to the target audience (using the "country" property in Wikidata) first, and generated rating second. Finally, we eliminated measurements for which the top suggestion for both audiences was the same, resulting in a set of 16 measurements.



(a) Area

(b) Height

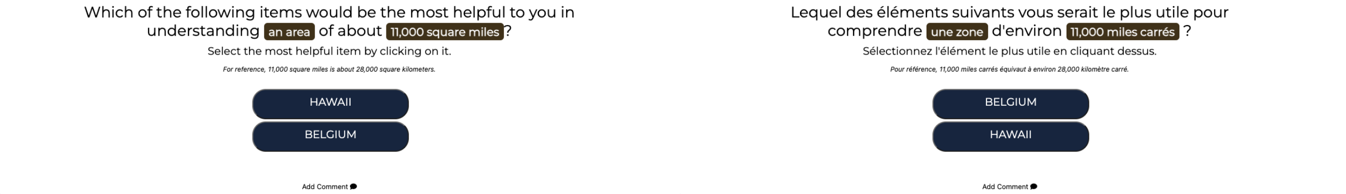
(c) Mass

(d) Length

(e) Population

Figure 3: Results of the measurement adaptability evaluation study, organized by measurement type. X axis shows model rating. Y axis shows human ratings. Color indicates object quality category based on model generated rankings. Overall, the model is able to distinguish low from medium and top quality objects, in accordance to human assessments.

A screenshot of the task can be found in Figure 4. Experiment text was translated in French for the French participants, and translations were validated by external proofreaders. To further ensure location truthfulness, we deployed additional measures, in addition to available qualifications for location and language. We disabled auto-translate and auto-select for all pages in the experiment, and asked participants to self identify whether they have lived at least 10 years, and are currently living, in the location they have been recruited from. We also asked participants to self report their local



Which of the following items would be the most helpful to you in understanding an area of about 11,000 square miles?

Select the most helpful item by clicking on it.

For reference, 11,000 square miles is about 28,000 square kilometers.

(a) US

Lequel des éléments suivants vous serait le plus utile pour comprendre une zone d'environ 11,000 miles carrés ?

Sélectionnez l'élément le plus utile en cliquant dessus.

Pour référence, 11,000 miles carrés équivaut à environ 28,000 kilomètre carré.

(b) FR

Figure 4: Interface for the two different audiences. Participants were asked to choose the most helpful reference object for understanding a given measurement, with each option representing the top ranked suggestion from our model for each audience. Left: United States, right: France

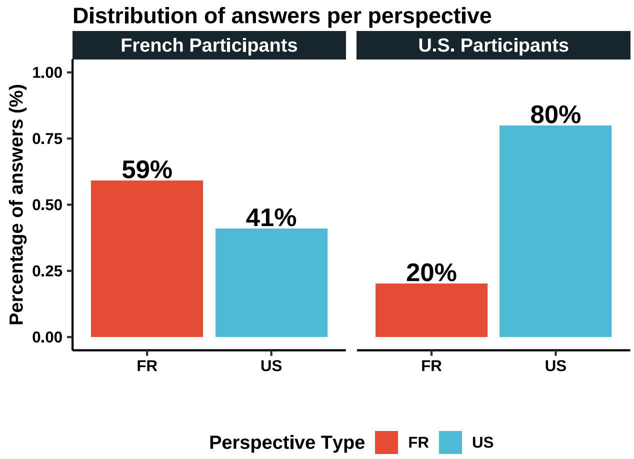


Figure 5: Participant preferences per audience.

time. Participants whose timezone did not align with the recruitment audience target, and whose responses revealed they were not currently living and/or have not lived at least 10 years in the target location were removed from the analysis. Both the order of tasks and the order of the two options within each task were randomly shuffled for each participant.

We pre-registered our analysis[10] prior to running our study. We recruited 150 participants in total, 75 US participants and 56 French participants through UHRS via Clickworker, and 19 French participants through Amazon Mechanical Turk. Participants were paid $2 for completing the study.

4.2.1 Results. A breakdown of responses by audience and perspective type can be found in Figure 5. We found that in both cases, participants significantly preferred the model suggestions that were tailored to their specific audience, with U.S. based participants choosing U.S. specific perspectives 80% of the time, and French participants choosing French perspectives 59% of the time.

We conducted a mixed effects logistic regression, with fixed effects for the location of the participants (whether they are non-U.S. or U.S.), and random effects for the participants and the questions, which revealed a statistically significant fixed effect of the audience location on the likelihood of choosing the non-U.S. option ($p < 0.001$). The estimated probability of non U.S participants choosing the non-U.S. perspective option, across all comparisons was 61.3% (SE = 0.0464, 95% CI [0.52, 0.70]) and the estimated probability of choosing the U.S. perspective option was 16.5% (SE = 0.0278, 95% CI [0.12, 0.23]).

Looking at the preference per specific measurement (Figure 6) offers additional insights into the effectiveness of different reference objects. For example, for a mass of 2,100 pounds, our model suggests an automobile manufactured by a U.S. company for U.S. audiences, and a French automobile for French audiences. Our study results reveal that participants strongly prefer the automobile that is closer to their region. We also observed the presence of references that may be useful across multiple audiences, such as using Hawaii for describing an area of 11,000 square miles. Hawaii is a globally known and popular travel destination, which is reflected in the percentage of responses for both audiences.

Of course it could be the case that while people prefer culturally tailored perspectives, such perspectives do not actually provide a sizeable boost in comprehension. To investigate this, we ran a followup study with 47 U.S.-based participants, testing their ability to estimate the quantities in Figure 6. We randomly assigned participants to see either all 16 of the U.S.-based references or all 16 of the France-based references and asked them to estimate each (e.g., the length of the Brooklyn Bridge in one condition or the length of the Pont de Normandie in the other). Figure 8 in Appendix A shows the result. U.S. participants who saw French reference objects were off by about a factor of 10 from the true values, whereas those who saw U.S. reference objects were only off by about a factor of 4. The latter is on par with best accuracy achieved in previous perspective-based estimation studies (see Figure 5 in [34] for comparison).

5 DISCUSSION

We summarize our findings in the following key takeaways:

An automated, low complexity, open-data model was empirically found to surface highly-rated perspectives: Overall, we find that we can effectively surface helpful analogies from large troves of encyclopedic information from open-source knowledge base systems such as Wikidata. Our feature evaluation revealed that a simple model using just Wikipedia information is just as powerful as including more complex and higher cost features, such as search volume and Google n-grams.

[10] https://aspredicted.org/c97ya.pdf

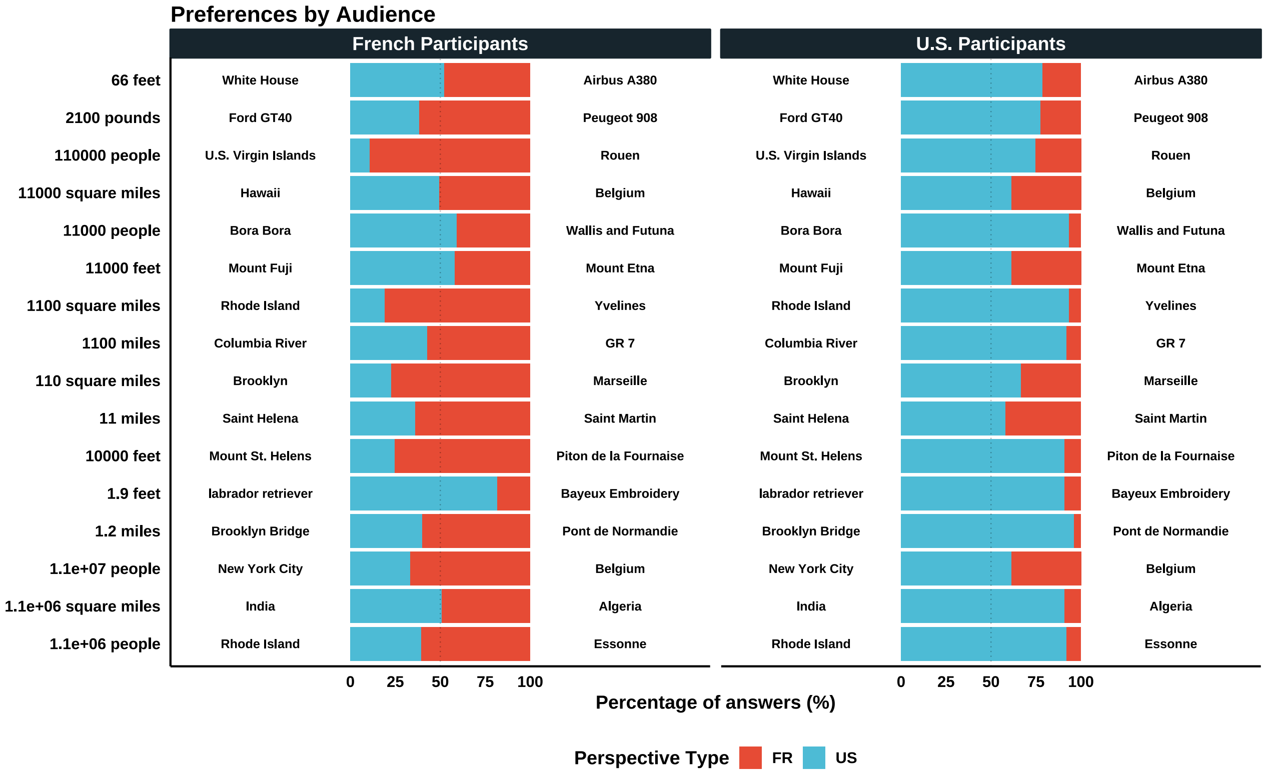


Figure 6: Measurement specific preferences by audience. For each presented measurement (Y axis), the breakdown of preferences for the two given options is depicted on the left for the French participants, and right for the U.S. based participants. Color denotes the audience adapted, model generated top ranked reference object, i.e. White House is the top ranked object when adapting to U.S. audiences and Airbus A380 is the top choice when adapting to French audiences.

The model can adapt to additional measurements without retraining: The model we developed in this work was relatively small, with training data on ratings of about just 300 objects, across 4 main attribute categories: area, height, length and mass. However, we found that our model is capable of generating rankings for additional dimensions, without requiring any retraining. In our evaluation study (Section 4), we found the highest performance in generating analogies for population, a category that was not present in the training data.

Wikipedia traffic can aid in creating more inclusive experiences: In addition to its ability to scale to different dimensions, we also evaluated our model's effectiveness towards scaling to different audiences. In our cross-audience study (Section 4.2), we found that participants significantly preferred the analogies that were tailored for their audience, highlighting the capacity of our model towards creating inclusive experiences. This result was achieved without requiring any retraining of the model; the sets of rankings for U.S. and France were both generated using the same base model, and simply adjusting the Wikipedia information according to each respective audience. Therefore, scaling to different audiences simply requires acquiring Wikipedia monthly views once for each audience of interest, and then using the same base model to re-rank candidate objects accordingly.

Existing knowledge bases need additional information to identify helpful reference objects: While our model achieved good performance overall, we still observed a relative struggle in ranking reference objects for length (Figure 3d). This result may be due to the fact that knowledge bases tend to contain large amounts of technical or obscure information that may not be useful for our application. For example, we found that Wikidata has records for many different small highways and roads that may not be useful for generating perspectives. On the contrary, measurements for more everyday objects, such as the length of a queen-sized bed, the weight of an orange, or the volume of a microwave, are most often not present in Wikidata despite being potentially useful for generating perspectives. Our findings highlight the potential for supplementing perspectives generated using our model with targeted crowdsourcing efforts to provide additional information.

5.1 Limitations & Future Work

In this work, we utilize a core characteristic of popular open-source knowledge bases such as Wikidata and Wikipedia, which is its multilinguality. With an active global network of contributors rigorously verifying and maintaining entries, we are able to access 322 currently active editions as of September 2023[11]. Nonetheless, there still lies a challenge in further discerning audiences. For example, English Wikipedia pageviews most likely contain traffic from people across the world and not just English speakers. In addition, it is currently not possible to distinguish pageview statistics at a country level, for countries whose native language is the same, that is, to distinguish pageviews for the U.S. versus the U.K., or Portugal versus Brazil. Therefore, additional audience information may be needed to ensure more fine-grained audience adaptability. Similar challenges are also identified in [39] and [37]. One such feature that we would like to explore is proximity to the user, similar to Kim et al. [19], for differentiating suggestions between countries whose first language is the same (e.g., surfacing Valencia for people in Spain and Chihuahua for people in Mexico when putting a population of 800,000 into perspective). Prior work has shown that geographically tailoring perspectives within a large country like the U.S. was not particularly effective relative to other strategies [34], however geographic tailoring between countries sharing a language merits investigation.

We evaluated the initial potential of our approach towards readily adapting to different audiences (Section 4.2) by examining the effectiveness of our approach on two separate countries, the United States of America and France and utilized two versions of Wikipedia with a sizeable difference in available information. There are about 59 million Wikipedia pages written in English, while French Wikipedia pages are about 13 million. Although our results highlight that Wikipedia information can be an effective proxy for helpfulness when generating audience-specific perspectives, the varying levels of availability of such information across different Wikipedia editions must also be considered. For example, at the time of writing, there were 13 million pages written in French, yet only about 660,000 in Greek. Lack of coverage in Wikipedia articles for some languages may lead to gaps in generating perspectives for different measurements for under-represented groups. In addition, even within an audience for which adequate Wikipedia data can be identified, it is important to consider the presence of different sub-populations, and how potential biases in the data could impact the model's suggestions. Potential future steps in addressing these issues could be: to incorporate targeted crowdsourcing within different audiences in order to recover culturally relevant items that may be missing, to present a range of perspectives instead of a singular item, and to present a variety of multipliers (e.g. twice the size of Lisbon to describe a population of 1 million for people in Portugal). The above challenges open up exciting future work in evaluating our approach in a coordinated effort spanning multiple countries and regions, to further assess performance, identify strengths and weaknesses, areas of improvement, whether a need for retraining or crowdsourcing arises and more.

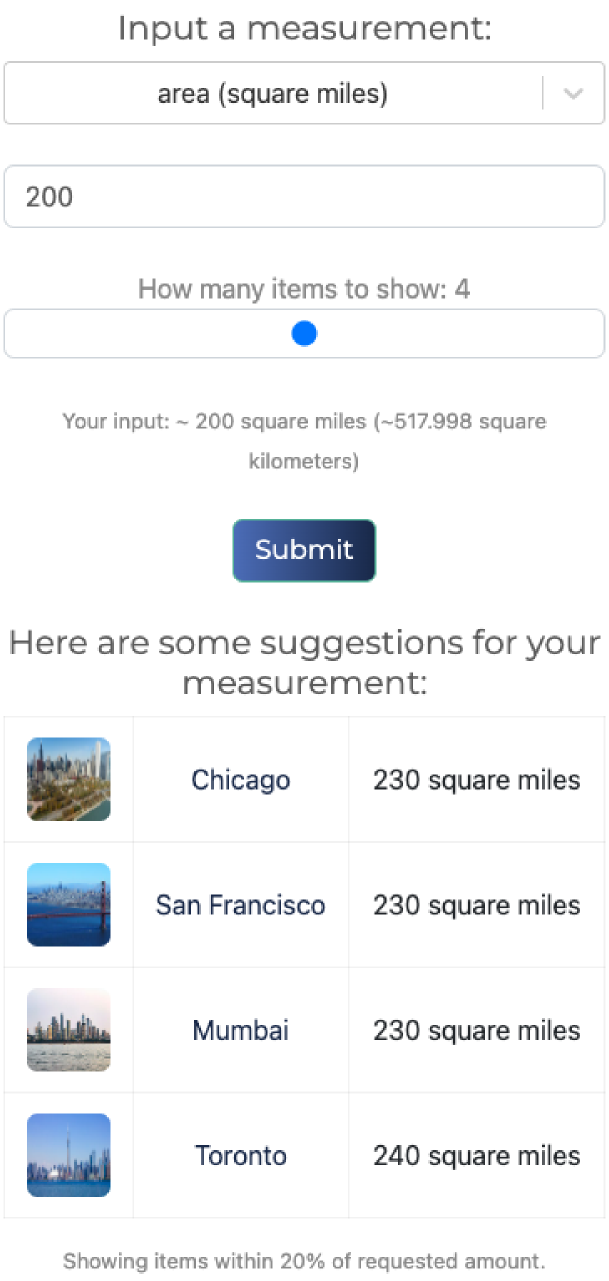
In our work, we employed an additional rank tie-breaking layer, by favoring reference objects with direct association to the audience we were adapting the rankings for. This was done by utilizing Wikidata's *country* property (P17) for each entity. However, we note that other types of region information that can be attached to certain objects may further improve adaptability.

A challenge in assessing the effectiveness of various features in a model that automatically generates perspectives lies in the subjective concepts of helpfulness and familiarity, which require collecting many human ratings for individual items. As our main goal was to develop a solution that automatically generates rankings and collecting human ratings for thousands of reference objects would be infeasible, we opted to explore the feasibility and effectiveness of models with a small training set. Indeed, we find that although it utilizes a small training set, our resulting model is able to adapt both to different measurement types and different audiences (Section 4). To reduce the number of training labels needed, we used a stratified sampling approach that is based on selecting items across different instances, therefore focusing on ensuring enough diversity of *types* of objects in our set, which could also offer some relative variety in proxies (e.g., countries may be generally more popular than roads). We are also interested in exploring the impact of other stratified sampling approaches, such as random selection, as well as increased training size on the effectiveness of our approach.

We note that the initially compiled database of candidate objects was not a exhaustive list of entities present in Wikidata, due to limitations in querying Wikidata. In addition, some level of pre-processing was required to further filter out entities that could not realistically be considered as candidates for use as perspectives. Such pre-processing was necessary due to the encyclopedic nature of knowledge base systems like Wikidata, that serve as central repositories of structured data, and are therefore organized in multiple layers. For example, there are 222,162 humans in Wikidata for whom there exists a height measurement, however, Barack Obama's height, which is highly likely to be top ranked by our model due to Wikipedia traffic, cannot realistically be considered a helpful perspective for 2 meters. However, this pre-processing was mainly conducted in considering a viable subset of reference objects for which collecting human assessments (a process that can be costly and time consuming) was meaningful, and is not required every time a new audience is considered.

Another limitation of utilizing knowledge bases such as Wikidata is that we are potentially missing common objects that are encountered in everyday life, which can be highly effective for generating perspectives for specific measurements. This absence is caused by the fact even though such objects exist in Wikidata, general measurements for them are not often recorded. Examples of such objects include things like a queen size mattresses (length), a microwave oven (volume), a school bus (capacity in people), a watermelon (weight) etc. These objects are however likely to surface via crowdsourcing methods, suggesting potential in supplementing wikidata data for broader coverage. To achieve this, a crowdsourcing task can be designed in the future in which participants can suggest items for various measurements and provide links to corresponding Wikipedia pages, making it possible for our model to generate rankings. In addition, our model does not currently adapt depending on the context surrounding a measurement, or depending on personalized preferences. Future work can explore combining different techniques that use context, audience, and personal preferences when generating tailored perspectives.

[11] https://meta.wikimedia.org/wiki/List_of_Wikipedias



Input a measurement:

area (square miles)

200

How many items to show: 4

Your input: ~ 200 square miles (~517.998 square kilometers)

Here are some suggestions for your measurement:

	Chicago	230 square miles
	San Francisco	230 square miles
	Mumbai	230 square miles
	Toronto	240 square miles

Showing items within 20% of requested amount.

Figure 7: Perspectives Tool. Given a specific measurement, e.g. area of 200 square miles, a list of the top ranked reference objects within 20% of the given measurement is returned.

As our primary aim was to build a model for automatically generating individual perspectives, we chose not to show multiple perspectives at once. However, we envision a system capable of combining multiple perspectives generated from our model. This opens up promising avenues for learning which combinations could lead to the most significant gains in comprehension and recall. Factors that could be explored include: the number of perspectives to show, the ratio of audience-tailored perspectives to global ones, and the number of multipliers to display. In this work we chose to use 1:1 perspectives, motivated by prior work that has found them to be the most effective [34]. However, multipliers of .5 or 2 were found to be second best, suggesting potential rewards for introducing them in combinations that expand the choice set for the user. We are currently working on an interactive tool for accessing a curated set of helpful perspectives for a variety of measurements (Figure 7).

5.2 Ethical Considerations

All human subjects studies involved in this work were reviewed and approved by the institutional ethics review board. Data sourced from the Wikidata open knowledge base to be used as candidates for perspectives are available under the Creative Commons CC0 license (public domain). We view our work as a promising step towards creating more inclusive experiences, as audience-aware perspectives can facilitate reasoning with numbers at a more global stage. We note that as our model uses Wikipedia traffic as a proxy of familiarity, it is possible for certain items that could be deemed not suitable to still be ranked higher over other items, for example specific weapons, necessitating an additional step of curation. We reviewed and excluded such cases in our pre-processing stages.

6 CONCLUSION

We present a model for automatically generating numerical perspectives; helpful analogies that can be used to re-express complex numerical information to make it easier to understand. Our Wikipedia-based model is able to generate comparable rankings to human assessments at a relatively low cost, compared to more computationally complex approaches like crowdsourcing. In addition, the model is able to successfully adapt suggestions for different audiences without incurring the costs of re-training. Our findings also highlight the importance of open sourced platforms like Wikipedia as a source of cross-cultural data.

Following the encouraging results identified in our evaluation studies, we undertook a more systematic collection of reference objects sourced from Wikidata across five measurement types (area, length, height, mass, population) and subsequently collected English Wikipedia information for all of them. Adapting this set to other audiences simply requires collecting Wikipedia information for the corresponding edition article, which can be easily retrieved from Wikidata using the Wikidata identifier. We release for public use and further research an updated dataset of over 400,000 Wikidata items with measurements and model ratings based on the English Wikipedia information. A curated subset will also be made available through our interactive tool.

In recent years there has been growing interest in how CHI can become more international and inclusive [38]. We hope this work will enable the global HCI research community to create tools that will make complex numerical information comprehensible and accessible to great numbers of people worldwide. This aligns with the field's emphasis on inclusivity and demonstrates how open source data and models can bridge cultural and linguistic barriers.

ACKNOWLEDGMENTS

We would like to thank all participants. We thank Jenny Wang for her help in developing and running the follow-up study on estimation error.

REFERENCES

[1] Elisa Barilli, Lucia Savadori, Stefania Pighin, Sara Bonalumi, Augusto Ferrari, Maurizio Ferrari, and Laura Cremonesi. 2010. From chance to choice: The use of a verbal analogy in the communication of risk. *Health, Risk & Society* 12, 6 (2010), 546–559.

[2] Pablo J Barrio, Daniel G Goldstein, and Jake M Hofman. 2016. Improving comprehension of numbers in the news. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2729–2739.

[3] Robbie Berg. 2013. Tropical cyclone report hurricane isaac (al092012) 21 august–1 september 2012. *National Hurricane Center Technical Report* (2013), 78.

[4] Michael Blastland and Andrew W Dilnot. 2009. *The numbers game: The commonsense guide to understanding numbers in the news, in politics, and in life.* Penguin.

[5] Christina Boyce-Jacino, Ellen Peters, Alison P. Galvani, and Gretchen B. Chapman. 2022. Large numbers cause magnitude neglect: The case of government expenditures. *Proceedings of the National Academy of Sciences* 119, 28 (2022), e2203037119. https://doi.org/10.1073/pnas.2203037119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2203037119

[6] Norman R Brown and Robert S Siegler. 1993. Metrics and mappings: a framework for understanding real-world quantitative estimation. *Psychological review* 100, 3 (1993), 511.

[7] Erick Burgueño Salas. 2022. Greece: Wildfire Area Burned 2022. https://www.statista.com/statistics/1264709/area-burned-by-wildfire-in-greece/, Last accessed on 2023-04-20.

[8] Yuru Cao, Hely Mehta, Ann E Norcross, Masahiko Taniguchi, and Jonathan S Lindsey. 2020. Analysis of Wikipedia pageviews to identify popular chemicals. In *Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications XII*, Vol. 11256. SPIE, 24–41.

[9] Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating Entity Disambiguation and the Role of Popularity in Retrieval-Based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4472–4485. https://doi.org/10.18653/v1/2021.acl-long.345

[10] Sunaina Dowray, Jonas J Swartz, Danielle Braxton, and Anthony J Viera. 2013. Potential effect of physical activity based menu labels on the calorie content of selected fast food meals. *Appetite* 62 (2013), 173–181.

[11] Yuya Fukano, Yosuke Tanaka, and Masashi Soga. 2020. Zoos and animated animals increase public interest in and support for threatened animals. *Science of the Total Environment* 704 (2020), 135352.

[12] Mirta Galesic and Rocio Garcia-Retamero. 2011. Communicating consequences of risky behaviors: Life expectancy versus risk of disease. *Patient education and counseling* 82, 1 (2011), 30–35.

[13] Mirta Galesic and Rocio Garcia-Retamero. 2013. Using analogies to communicate information about health risks. *Applied Cognitive Psychology* 27, 1 (2013), 33–42.

[14] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 89–101.

[15] Brent Jaron Hecht. 2013. *The mining and application of diverse cultural perspectives in user-generated content.* Ph. D. Dissertation. Northwestern University.

[16] Jessica Hullman, Yea-Seul Kim, Francis Nguyen, Lauren Speers, and Maneesh Agrawala. 2018. Improving comprehension of measurements using concrete re-expression strategies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[17] Brian Kernighan. 2019. *Millions, Billions, Zillions: Defending Yourself in a World of Too Many Numbers.* Princeton University Press.

[18] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting scientific results in perspective: Improving the communication of standardized effect sizes. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[19] Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating personalized spatial analogies for distances and areas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 38–48.

[20] Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. 2018. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 659–664.

[21] Richard P Larrick and Jack B Soll. 2008. The MPG illusion. *Science* 320, 5883 (2008), 1593–1594.

[22] Benjamin Lee, Dave Brown, Bongshin Lee, Christophe Hurter, Steven Drucker, and Tim Dwyer. 2020. Data visceralization: Enabling deeper understanding of data using virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1095–1105.

[23] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 618–626. https://doi.org/10.1145/3289600.3291021

[24] Yinghong Ma, Jiaoyang He, and Qinglin Yu. 2019. Modeling on social popularity and achievement: A case study on table tennis. *Physica A: Statistical Mechanics and its Applications* 524 (2019), 235–245.

[25] Michael A Martin. 2003. "It's like... you know": The use of analogies and heuristics in teaching introductory statistical methods. *Journal of Statistics Education* 11, 2 (2003).

[26] Katie Meehan, Jason R. Jurjevich, Nicholas M. J. W. Chun, and Justin Sherrill. 2020. Geographies of Insecure Water Access and the Housing–Water Nexus in US Cities. *Proceedings of the National Academy of Sciences* 117, 46 (Nov. 2020), 28700–28707. https://doi.org/10.1073/pnas.2007361117

[27] Marc Miquel-Ribé and David Laniado. 2019. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 620–629.

[28] John C Mittermeier, Uri Roll, Thomas J Matthews, Ricardo Correia, and Rich Grenyer. 2021. Birds that are more commonly encountered in the wild attract higher public interest online. *Conservation Science and Practice* 3, 5 (2021), e340.

[29] Vikram Mohanty, Alexandre L. S. Filipowicz, Nayeli Suseth Bravo, Scott Carter, and David A. Shamma. 2023. Save A Tree or 6 Kg of CO2? Understanding Effective Carbon Footprint Interventions for Eco-Friendly Vehicular Choices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 238. https://doi.org/10.1145/3544548.3580675

[30] John Allen Paulos. 1988. *Innumeracy: Mathematical illiteracy and its consequences.* Macmillan.

[31] Xiaoli Qiao and Jessica Hullman. 2018. Translating scientific graphics for public audiences. In *Proceedings of the VisGuides: 2nd Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization. IEEE VIS.*

[32] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. 2010. Characterizing and modeling the dynamics of online popularity. *Physical review letters* 105, 15 (2010), 158701.

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[34] Christopher Riederer, Jake M Hofman, and Daniel G Goldstein. 2018. To put that in perspective: Generating analogies that make numbers easier to understand. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[35] Jesus San-Miguel-Ayanz, Tracy Durrant, Roberto Boca, Pieralberto Maianti, Giorgio Liberta', VIVANCOS Tomas Artes, FELIX OOM Duarte Jacome, Alfredo Branco, RIGO Daniele De, Davide Ferrari, Hans Pfeiffer, Rosana Grecchi, and Daniel Nuijten. 2022. Advance Report on Wildfires in Europe, Middle East and North Africa 2021. https://publications.jrc.ec.europa.eu/repository/handle/JRC128678. https://doi.org/10.2760/039729

[36] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1591–1600. https://doi.org/10.1145/3038912.3052716

[37] Benjamin K Smith and Abel Gustafson. 2017. Using wikipedia to predict election outcomes: online behavior as a predictor of voting. *Public Opinion Quarterly* 81, 3 (2017), 714–735.

[38] Angelika Strohmayer, Cayley MacArthur, Velvet Spors, Michael Muller, Morgan Vigil-Hayes, and Ebtisam Alabdulqader. 2019. CHInclusion: Working toward a more inclusive HCI community. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–10.

[39] Michele Tizzoni, André Panisson, Daniela Paolotti, and Ciro Cattuto. 2020. The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic. *PLoS computational biology* 16, 3 (2020), e1007633.

[40] Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: predicting book sales before publication. *EPJ Data Science* 8, 1 (2019), 1–20.

A ESTIMATING AUDIENCE SPECIFIC PERSPECTIVES

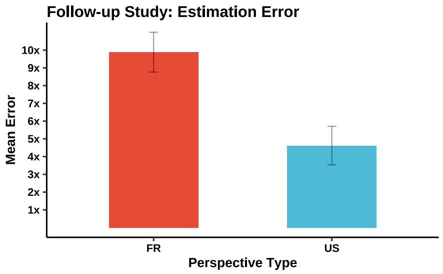


Figure 8: Results from the follow-up study on estimating quantities described in Section 4.2. In the study, 47 online participants were instructed that they would be estimating quantities that they may or may not be familiar with. All the participants were based in the U.S. and were randomly assigned to estimate the measurements of either the U.S. or French reference objects found in the 16 rows of Figure 6. Participants were shown a set of possible answers that spanned the ground truth and chose the one they believed to be closest to it. Answer choices for a given row in the table were the same across conditions and the ground truth values are roughly equal across conditions as well. The order of the questions was randomized for each participant. We compute error as follows, adopting the approach of [6, 34]. For each estimate, we take the absolute value of log_{10}(estimate) - log_{10}(ground truth), then take the mean of these values by condition. Finally we raise 10 to the power of the mean of each condition. This error metric expresses the multiple by which the average estimate is incorrect. Error bars are +/- 1 standard error. Participants who saw French reference objects were off by about a factor of 10 from the true values, whereas those who saw U.S. reference objects were only off by about a factor of 4, which is on par with prior estimation accuracy studies.