Representing data in words: A context engineering approach

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated remarkable potential across a broad range of applications. However, producing reliable text that faithfully represents data remains a challenge. While prior work has shown that task-specific conditioning through in-context learning and knowledge augmentation can improve performance, LLMs continue to struggle with interpreting and reasoning about numerical data. To address this, we introduce wordalisations, a methodology for generating stylistically natural narratives from data. Much like how visualisations display numerical data in a way that is easy to digest, wordalisations abstract data insights into descriptive texts. To illustrate the method's versatility, we apply it to three application areas: scouting football players, personality tests, and international survey data. Due to the absence of standardized benchmarks for this specific task, we conduct LLM-as-a-judge and human-as-a-judge evaluations to assess accuracy across the three applications. We found that the wordalisation methods reduces misrepresentation of the data and shows the potential to improve produce engaging communication about data. We further describe best practice methods for open and transparent development of communication about data.

1 Introduction

Data science has become an important tool in business (Davenport, 2020), biology (Greener et al., 2022), epidemiology (Latif et al., 2020), journalism (Cairo, 2016) and beyond. One way in which data science complements, and in some cases goes beyond, machine learning is in the use of visualisation to better understand datasets. Visualisation helps machine learning engineers both to understand their models and to communicate findings to non-technical users (Zhang et al., 2020). Good visualisation practice involves understanding how to effectively communicate key facts about data without misleading the audience. For example, a series of articles by data journalist Lisa Charlotte Muth, and more recently Rose Mintzer-Sweeney, illustrate how careful design choices — around colours, fonts, the placement of text and choice of chart — are required when creating visualisations so as to convey information as clearly as possible (Mintzer-Sweeney and Muth, 2024). Visualising data is not a neutral process, but rather an outcome of numerous decisions about how we want to present it to an audience (Kennedy et al., 2016).

Large language models (LLMs) currently demonstrate impressive mathematical and reasoning capabilities. However, despite their strong problem-solving strategies, they often struggle with presenting numerical data, partly because they treat numbers as text tokens rather than as quantitative information (Yang et al., 2025; Li et al., 2025). In this article, we focus on communicating data in words using LLMs. Our aim is to "wordalise" data in a reliable way, going from raw data to a statistical model, to a normative model (i.e. an interpretation of the numerical value) and finally a wordalisation (see Fig. 2 for an example of these steps). The wordalisation (a term we coin here) is generated by an LLM using a prompt constructed according to four task-agnostic steps, described in detail in Section 3, which build on in-context learning and few-shot prompting examples (Brown et al., 2020). We have built a prototype application to demonstrate our idea, available at `https://wordalisations.streamlit.app/`, which we encourage readers to visit to better understand our aims.

Unlike data-to-text generation (Lin et al., 2024; Puduppully et al., 2022; Osuji et al., 2024), which aims to produce text from structured data such as tables using neural networks (Iso et al., 2019), we rely on a trained LLM to generate text (Wiseman et al., 2019). Our work is under the umbrella of data-driven storytelling (Schröder et al., 2023; Shao et al., 2024; Zhao and Elmqvist, 2023), which

(a) Workflow: A statistical model is applied to the raw data. The statistical description is converted into a synthetic text using a normative model. Combined with an appropriate prompt, the model generate a wordalisation, a high-quality, easy-to-digest text that faithfully represents the data.

Statistical model:

Factor	Z-score
Non-penalty expected goals	2.101
Final third passes	-0.663
Final third receptions	-0.716
Ground duels	-0.142
Air duels	-0.487
Smart passes	-0.028
Goals	2.041
Assists	-0.717
Key passes	-0.474
Minutes	1.873

Normative model: Here is a statistical description of [H. Kane], who played for [3201] minutes as a Forward. He was [outstanding] in [non-penalty expected goals] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [below average] in [final third passes] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [below average] in [final third receptions] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [average] in [ground duels won] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [average] in [air duels won] adjusted for possession per 90 minutes compared to other players in the same playing position. He was average in [smart passes] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [outstanding] in [goals] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [below average] in [assists] adjusted for possession per 90 minutes compared to other players in the same playing position. He was [average] in [key passes] adjusted for possession per 90 minutes compared to other players in the same playing position.

Wordalisation: Harry Kane is an exceptional goal scorer who consistently creates high-quality scoring opportunities for himself. His outstanding goal and expected goals numbers highlight his clinical finishing ability and knack for finding dangerous positions. However, his contributions in creating chances for teammates are limited, as evidenced by his below-average assist and key pass numbers, and his involvement in the final third is also relatively low. Compared to other forwards, Kane's prolific goal-scoring ability makes him stand out significantly, despite the less-impressive numbers in other metrics.

(b) Example wordalisation. The statistical model, here the z-score, of the football player Harry Kane is shown as a table (top). These are converted into a synthetic text using a normative model (middle). The normative model, in combination with the prompting template described in Section 3, is used to generate the example wordalisation for Harry Kane (bottom).

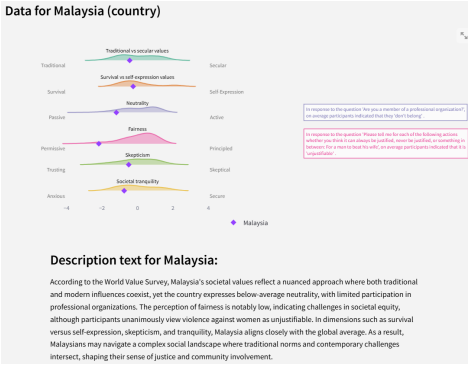
Figure 1: Workflow and example wordalisation.

aims to turn raw data into easy-to-read and easy-to-understand plain stories to help the user gain insights into data. In contrast to automatic captioning methods (You et al., 2016; Hossain et al., 2019; Herdade et al., 2020), wordalisations aim to generate text directly from data, in an analogous way to how visualisation turns data into graphs. Visual elements can be used to support the understanding of the text, as shown in our demo application.

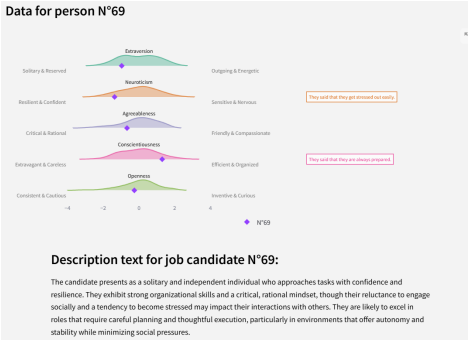
Using prompt engineering to embed specific knowledge into LLMs (Ahmed et al., 2024; Liu et al., 2025; Song et al., 2025) is quickly becoming the go-to method for improving answers provided by LLMs and counteracting hallucinations (Chen et al., 2024). Further, prompt engineering (Amatriain, 2024; Schulhoff et al., 2024; Liu et al., 2023) can be used to shape the wordalisation generated for the specific user application. An advantage of using prompt engineering over model fine-tuning is that evaluating different prompts has a much lower computational cost, performing as well and in come cases better than fine tuned models (Liu et al., 2022). As Andrej Karpathy recently pointed out, what is needed goes further than prompt engineering, to encompass the idea of "context engineering" to provide language models with the right information, tools, and instructions appropriately formatted to enable reliable task completion (Blog).

2 Application and Workflow

The originality of our work lies in the way we use prompt engineering applied to the LLM to convert raw numerical data into explanatory text that returns information consistent with the data itself, but also provides a judgment about the individual. The design aim of the wordalisation is to be both engaging to the reader and an accurate representation of the data source. Wordalisation can be applied across domains and to show this, we apply our framework to three application areas: football scouting, personality test analysis and international survey data exploration. Our method is applicable in scenarios where the user would like to compare a specific entity (single item in the dataset) to the larger population (the full dataset). We have built a Streamlit application (Richards, 2023) to demonstrate our idea, and made it available here `https://wordalisations.streamlit.app`. Figure2 gives examples of visualisations and corresponding wordalisation for each application area.



(a) Data visualisation and wordalisation of the World Values Survey Wave 7: Case study of Malaysia.



(b) Data visualisation and wordalisation of the personality test: Case study of the person 69.

Figure 2: Data visualisation and wordalisation for psychometric and sociodemographic datasets: Applications to World Values Survey and personality assessment results.

Instead of simply telling the reader values or rankings, the wordalisation engages the reader. For example, by using familiar footballing terminology, while simultaneously accurately portraying the player's skills in shooting, passing the ball, competing defensively, heading and other skills an overview of the player is created.

In the demo application, we also integrate the wordalisation with a chat functionality, using retrieval-augmented generation (RAG) (Lewis et al., 2020), allowing the user to ask questions about a specific individual or entity, in our case player, person, or country.

The overall workflow for building a wordalisation is illustrated in Fig. 2). Initially, informative statistics are calculated from the data, after which a normative model converts these numbers into a statistical description. This automated generation is done without the use of an LLM, using a template. As a result, it generates bland formulaic, but accurate texts. This synthetic text is then passed to a large language model as pat of the prompt we outline below in order to generate the final wordalisation.

3 Constructing the prompt

We now give a step-by-step account of how to construct the prompt underlying a wordalisation in four steps:

Tell it who it is. The first step is to set the system prompt, which defines the role of the language model that will be used to generate text.

Tell it what it knows. Here we give user-assistant prompts in the form of question-answer pairs that provide both information on how the language model should answer questions and provide background knowledge.

Tell it what data to use. This step converts the content of the dataset and a selected data point into words. It allows for converting the statistical model into the normative model.

Tell it how to answer. We provide examples of how the language model should write the final description from the text generated at the "tell it what data to use" stage.

We now detail the 4 steps and give an example for each. By using the appropriate expander bar in the Streamlit app it is possible to see how the steps are implemented for each individual entity.

3.1 Tell it who it is

When calling an application programming interface (API) of a large language model, the system prompt tells the LLM the role the assistant should take when answering the questions. In the case of the football scout, we write:

> "role": "system",
> "content": ("You are a UK-based football scout. You provide succinct and to the point explanations about football players using data. You use the information given to you from the data and answers to user/assistant pairs to give summaries of players.")

In generating a response the LLM pays special attention to this first system prompt, so it is important to tell it clearly what to do. For wordalisation we suggest being specific about the type of task that will be carried out, but to not yet give details of how it will be performed.

3.2 Tell it what it knows

After the system prompt we provide a sequence of question-answer (Q&A) pairs, which help the LLM interpret the data it will be provided. These Q&A pairs primarily illustrate how a domain expert would explain their knowledge to a non-expert. It is therefore essential that these answers both clarify what the data can tell us and highlight its limitations. These can be written by an expert and/or adapted from a reliable source. For example, for personality traits we used descriptions from Wikipedia of the big five personality traits (Matthews et al., 2003) and for the football scout application we asked a domain expert to describe football statistics in terms of player performance.

Question	Answer
What is a forward?	A forward is the football player whose primary role is to score goals or to assist other players in scoring goals. They often find themselves receiving the ball in the box or making runs to receive the ball high up the pitch.
What does it imply if a forward is good or excellent in terms of goals?	A forward who is good or excellent in terms of goals scores a lot. Goals are the most important thing in football, but we need to check that the player also does well in expected goals in order to check whether they might just have been lucky.

Table 1: Example of a question annd answer pair for the football scout example.

Constructing Q&A pairs by hand ensures the relevance and accuracy of the Q&A pairs. However, this task could also be semi-automated using an LLM. For such an approach, it is important to ensure that all phrases from the original text are accurately represented. Q&A pairs should give insight into the data over and above simple definitions, and convey the desired language and tone.

3.3 Tell it what data to use

Wordalisation assumes an underlying *statistical model* to effectively interpret data. In the case of the three applications presented here, we assume we can assess an individual (or entity) i in terms of the z-score over one or more metrics,

$$z_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j},$$

where μ_j is the mean and σ_j the standard deviation of the metric j. The z-scores are useful because they standardise the metrics and allow us to, for example, compare a player in terms of receiving and passing the ball on the same scale, defined by how far their z-score is from that of the average player in terms of standard deviations. Following statistical convention, one standard deviation can be considered to be larger than usual, two standard deviations can be considered to be significantly larger, and so on.

The z-scores, in our case, are then formatted as a synthetic text, with phrases based on threshold ranges. For the football scout application, the threshold for z-scores are chosen as follows: above 1.5 is described as "outstanding", between 1.0 and 1.5 is described as "excellent"; between 0.5 and 1.0 is described as "good"; between -0.5 and 0.5 is described as "average"; between -1.0 and -0.5 is described as "below average"; and below -1.0 is described as "poor". We apply these thresholds to all the metrics for the player to generate a synthetic text. An example of this for Harry Kane is shown in Fig. 2. For illustration purposes, square brackets are used to highlight the selected phrases, and all square brackets are omitted in the final prompt. In these texts, we avoid presenting specific numbers by instead presenting a written interpretation of the data. We can think of this as a *normative model* for how we interpret the player's actions: more goals are better, as are more passes into the final third, etc. It is important to note that a wide range of statistical models can be applied here.

3.4 Tell it how to answer

The "tell it how to answer" step is a two-part process that involves providing the LLM with guidelines on response structure and incorporating a few-shot prompting technique. Firstly, the instruction part of this step can be used to detail instructions such as "answer in 3 concise sentences" or "only use the data provided". These directives serve as a precise framework, specifying the desired format and tone of the LLM's responses.

"role": "user",
"content": (Please use the statistical description enclosed with "‘ to give a concise, 4 sentence summary of the player's playing style, strengths and weaknesses. " "The first sentence should use varied language to give an overview of the player. " "The second sentence should describe the player's specific strengths based on the metrics. " "The third sentence should describe aspects in which the player is average and/or weak based on the statistics. " "Finally, summarise exactly how the player compares to others in the same position.)

Secondly, few-shot prompting is used to demonstrate examples of how the LLM should adhere to the instructions. This involves including example pairs of statistical descriptions and corresponding assistant responses within the prompt. These pairs also reinforce the explicit directive given. We found this step to be essential, significantly improving the quality of the LLM's responses. These examples influence various aspects of the response, including structure, formatting, tone, tense, and which elements to emphasise. The LLM then closely adheres to these patterns when discussing similar data, ensuring consistency and clarity in the responses.

4 Evaluation and documentation

Assessing the performance of a large language model on a specialised task, cannot be done via a standardised benchmark because each wordalisation application has its own unique and subjective set of goals and assessment criteria (Raji et al., 2021; Kaddour et al., 2023; Vendrow et al., 2024). Instead of benchmarking we propose a holistic approach for evaluating the performance using a combination of LLM-as-a-judge (Gu et al., 2024; Li et al., 2024) and human-as-a-judge methods with thoughtful documentation.

4.1 LLM-as-a-judge

To measure the accuracy, we generate texts in three different ways: as a *control* comparison, we omit the synthetic text describing the data from the prompt. As a *numerical only* comparison we send only the raw statistics. We compare these two approaches to the *full* wordalisation. The resulting texts from these three methods were evaluated by asking Google's "gemini-2.0-flash": "Does the generated text contain implications or draw conclusions that are not supported by the statistical description?" which gave a binary classification of 'yes' or 'no'. To discourage the LLM from declining to respond when no description is given, we added the sentence "If no data is provided, answer anyway, using your prior statistical knowledge".

The control prompt generates factual inaccuracies in 91% of cases, the numerical statistical description gave inaccuracies in 50% of cases, while using the full Wordalisation texts and 18%. Pairwise comparisons using z-tests for proportions confirmed that using the normative description of the data significantly increases accuracy in the text compared to both the control and numerical conditions ($p < 0.001$ in both cases).

To investigate whether it is possible to accurately reconstruct the statistical description of a data point from the wordalisation we compare the statistical description that lies in the *synthetic text* to a *reconstructed synthetic text*. Given a 'wordalisation', we instruct the LLM to classify each factor, according to the classes determined by the normative model, then to reconstruct the synthetic text. The original synthetic text serves as the ground truth for evaluation. The accuracy of the reconstructed statistical description serves as a measure of how faithfully the wordalisation represents the data point. For example, for the wordalisation of Harry Kane shown in Fig. 2, it is possible to (correctly) predict the class 'outstanding' for the 'goals' factor. We prompt the LLM to predict a classification for each relevant factor. We then compare the 'true' class determined by data and the normative model with the predicted class. In cases where the generated json file could not be parsed, the data was discarded. To take into account random variations in the wordalisation due to the stochastic nature of the language model, we passed each prompt to the LLM multiple times, generating several wordalisations for each data point.

For the personality test and international survey applications, the high accuracy scores achieved when the normative model was used, shown in blue in Fig. 3, indicate that the wordalisation were very faithful to the statistical descriptions provided by the normative model. For all applications, the test accuracies, where a description according to the normative model was provided, were higher across all factors compared to the control, where statistical data was omitted. This performance difference

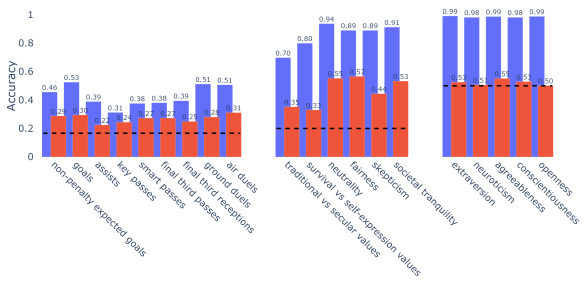


Figure 3: Comparison of the class labels generated by the normative model with classes reconstructed from the wordalisation. For each application, football scout (left), international survey (middle) and personality test (right), multiple wordalisations were generated for each data point, so that at least 10 valid reconstructions per data point were found, and the mean accuracy is taken over all wordalistations. We compare the accuracy of the model for two different prompts, one in which data in the form of synthetic text, generate according to the normative model, was given (blue) and one in which the data was omitted (red). The dashed line indicates the expected accuracy if the class labels were randomly chosen according to a uniform probability distribution and lie at an accuracy of $\frac{1}{6}$, $\frac{1}{5}$ and $\frac{1}{2}$ for each application respectively.

is further evidence for the influence of the statistical description generated by the normative model on the wordalisation. For comparison, we also indicate, using dashed lines in Fig. 3, the accuracy expected in the case where each factor label was guessed, that is, selected at random with each label having equal probability of selection. For the football scout and international survey applications, the performance of the control was better than guessing across all factors, while the performance of the control for the personality test application was comparable to guessing. A likely explanation for this performance difference is the presence of relevant information in the LLM training data, which could plausibly be the case for football players and countries but not for anonymised individuals who took a personality test.

Fig. 3 also shows that the overall performance of the football scout application when a synthetic text description given is lower than that of the two other applications. We postulate that this is a result of prompting, since in the football scout application the LLM was encouraged to generate more creative texts, while in the personality test and international survey applications the LLM was encouraged to respond in a more formulaic style. For this reason it was more difficult for the LLM to reconstruct the statistical descriptions of a football player from a wordalisation, compared to a person or a country. Depending on the use case and intended audience of the application this is of more or less concern. For example, a football expert might favour an engaging, descriptive text, while a human resources manager or social scientist might prefer an unadorned, to-the-point summary.

4.2 Human-as-a-judge

For human evaluation we compared the *wordalisation* to two methods: a *control* where the model is asked to describe an entity without access to numerical data. In the international survey and football scout example, the LLM used its training data to write descriptions about the country or football player, while for personality tests no fall back data was available. For the *statistical* method, the model is provided with z-scores and asked to generate a description based on these values. Using these three methods, we generate description for 30 entities - 10 from each example-resulting in total of 90 texts. Human raters were recruited through social media posts to go to an app (`https://wordalisations-evaluation.streamlit.app/`) where they were presented with a single entity at a time, along with its corresponding visualisation and a description from either the control, the statistical or the wordalisation treatment. To prevent direct comparisons and learning effects, each rater saw only one strategy per entity and never evaluates multiple descriptions of the same entity. Raters may evaluate as many descriptions as they wish and may stop at any point. For each description, we asked raters to assess the faithfulness of the description to the visualisation; its engagement; its usefulness for understanding or decision-making; and whether it contains hallucinations, defined as unsupported or fabricated claims. Responses are collected using ordinal Likert-style scales, with optional free-text comments. The evaluation interface presents one description per page, and a new entity–method pair is sampled immediately after submission. No personally identifiable information is collected; participants are informed of the study purpose and their right to withdraw at any time. We had 121 responses (from 26 unique individuals) with 42 texts from the control prompt, 40 texts from the statistical prompt and 40 texts form the wordalisation prompt.

Fig. 4 shows how the wordalisation and statistical vastly texts outperformed the statistical text in

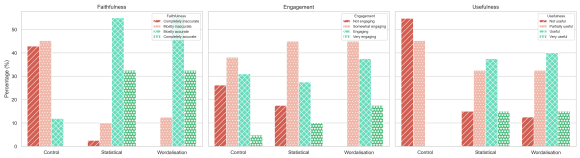


Figure 4: Human evaluation of three text variants (control, statistical, and wordalisation) across three metrics: faithfulness, engagement, and usefulness. Percentages of categorical ratings are shown for each text variant.

terms of faithfulness and usefulness ($\chi^2 = 71.00$ for faithfulness and $\chi^2 = 40.50$ for engagement, both $p < 0.001$). Engagement was also significantly different when compared across all methods ($\chi^2 = 13.93$, $p = 0.03$). Importantly, when comparing only the statistical and wordalisation texts, we see that while they do not differ in terms of faithfulness ($\chi^2 = 1.11$, $p = 0.77$) or usefulness ($\chi^2 = 0.12$, $p = 0.99$), there was a significant difference in engagement ($\chi^2 = 8.43$, $p = 0.04$), with wordalisation preferred by the participants over statistical descriptions.

4.3 Documentation

Following the model cards framework (Mitchell et al., 2019; Crisan et al., 2022; Gebru et al., 2021; Crawford, 2021), we emphasise the importance of clearly stating the model we are imposing on the data when creating the wordalisation: detailing how numerical values are translated into words, incorporating background information into prompts for the LLMs, and documenting the limitations of the wordalisation. This approach provides transparency on the risks of applying wordalisation to a new domain without performing extensive quantitative analysis. Model cards offer a flexible framework for innovative thinking around best practices in modelling (Tang et al., 2024). We documented each of our applications in the model card style adapted from (Mitchell et al., 2019) and can be found here[1].

All code developed for this work has been made publicly available in a public repository here[2]. This includes the full implementation for the 3 applications and the evaluation pipeline used for assessment. By releasing the application and evaluation code, we aim to support transparency, reproducibility and broader adoption of the methodology.

[1] https://anonymous.4open.science/r/wordalisation-4088/model%20cards/model-cards.md

[2] https://anonymous.4open.science/r/wordalisation-4088/

5 Discussion

The key to the wordalisation approach is the combination of both statistical and normative models to represent data in words. It is this which leads us to our four steps of telling an LLM 'who it is', 'what it knows', 'what data to use' and 'how to answer' as a way to produce convincing texts. The 'what data to use' step defines the statistical model of how we interpret the numbers in terms of z-scores and then a normative model provides our interpretation of these values. The other three steps give a text-based model of how we should interpret those scores, giving further background to our normative model. While z-scores are a relatively simple statistical model, we envisage (and are currently implementing) more advanced statistical models — such as regression, principal component analysis and moving averages — as the basis for wordalisation.

Instead of fine-tuning a model, we force the pretrained LLM to produce the texts which reflect our own model of that data. Previous work has shown that extensive fine-tuning of LLMs often does not significantly outperform (and in some cases underperforms) methods based on prefixed prompts on benchmark tasks (Li and Liang, 2021; Jia et al., 2022). Indeed, prompt-based or in-context learning (Radford et al., 2019; Liu et al., 2023; Brown et al., 2020), of which our wordalisations are examples, are a powerful way of utilising the power of LLMs in a way which respects their limitations. Our four-step automated method underlines just how well prompting can perform in application areas.

In human evaluations, we found that wordalisations outperformed a purely statistical description of the data in terms of engagement, while maintaining faithful and useful descriptions of the underlying data. This result contrasted with LLM evaluations, where the wordalisations for the football scout application, in particular, were unable to reconstruct the exact normative description. What human evaluation identified is that rather than containing inaccuracies, they simply contained more varied language and were less formulaic. We thus see our human evaluations as highlighting the limitation of the LLM-as-a-judge approach, rather than as a benchmark test of performance.

Ultimately, we do not envisage quantitative testing, by humans or machines, as the way to assess wordalisations. The wordalisation task involves

generating text from structured data, making it a fundamentally different objective than captured by established benchmarks for Natural Language Processing (NLP) tasks. For this reason, we have also adopted an approach that builds on model cards (Mitchell et al., 2019), emphasising the quality of documentation of the normative and statistical models we use, and the consequences of using these models, over output testing. As language models become integrated into an increasing number of tools, it is essential to provide transparent documentation of the thought process behind the prompt engineering (Liu et al., 2023). In future work, we will give a more complete outline of the model card approach to prompting.

Wordalisations, similar to visualisations, are powerful tools for data communication. They have the potential to make all types of data more accessible and comprehensible to a wider audience. However, evaluating them is inherently subjective, as absolute accuracy is often unattainable (Tong and Zuo, 2021; Smith, 2023). In our work, we have embraced this reality by prioritising transparency and honesty, akin to practices in visualisation. Rather than solely emphasising numerical evaluations, we advocate for providing thoughtful documentation through model cards and communicating model capabilities and limitations. Benchmarking and evaluation, while having some value, should support this process rather than be the sole target. We can compare this process with journalism: journalists rarely focus on benchmarking their work, but attempt to maintain high standards (Mintzer-Sweeney and Muth, 2024; Cairo, 2016). It is also consistent with the idea of context engineering (Blog): we are no longer at the stage of benchmark accuracy tests, but looking for ways to develop best practices for working with LLMs. Those exploring prompting should focus on documenting and standardising their methodologies to build a robust framework for future use. Looking ahead, prompting will play an increasingly significant role in software development and having a clear structure for documentation will be essential.

Limitations

A current limitation of the wordalisation method is its exclusive reliance on raw numerical data; it does not yet support the integration of categorical variables. The examples presented is intended as an illustration of the methods. The primary use case of this wordalisation is educational. It shows how to convert a dataframe of raw data into a text. A secondary use case might be for users to understand more about the Big Five personality test, World Value Survey, or Football player datasets. This version is not suitable for professional use, such as by a recruiter. However, the methods used here do have the potential to be adopted to a professional setting.

References

Awais Ahmed, Mengshu Hou, Rui Xi, Xiaoyang Zeng, and Syed Attique Shah. 2024. Prompt-eng: Healthcare prompt engineering: Revolutionizing healthcare applications with precision prompts. In *Companion Proceedings of the ACM Web Conference 2024*, New York, NY, USA. Association for Computing Machinery.

Xavier Amatriain. 2024. Prompt design and engineering: Introduction and advanced methods. *Preprint*, arXiv:2401.14423.

Langchain Blog. The rise of context engineering. https://blog.langchain.com/the-rise-of-context-engineering/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alberto Cairo. 2016. *The truthful art: Data, charts, and maps for communication*. New Riders.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2024. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439.

Thomas Davenport. 2020. Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review*, 2(2):5.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Joe G Greener, Shaun M Kandathil, Lewis Moffat, and David T Jones. 2022. A guide to machine learning for biologists. *Nature reviews Molecular cell biology*, 23(1):40–55.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2020. Image captioning: Transforming objects into words. *Preprint*, arXiv:1906.05963.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6).

Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, and Hiroya Takamura. 2019. Learning to select, track, and generate for data-to-text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 2102–2113. Association for Computational Linguistics.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Helen Kennedy, Rosemary Lucy Hill, Giorgia Aiello, and William Allen. 2016. The work that visualisation conventions do. *Information, Communication & Society*, 19(6):715–735.

Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, Adrian Weller, and 1 others. 2020. Leveraging data science to combat covid-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 1(1):85–103.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Haoyang Li, Xuejia Chen, Zhanchao XU, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Qing Li, and Lei Chen. 2025. Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models. *Preprint*, arXiv:2502.11075.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2024. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Haoxin Liu, Wenli Zhang, Jiaheng Xie, Buomsoo Kim, Zhu Zhang, Yidong Chai, and Sudha Ram. 2025. Few-shot learning for mental disorder detection: A continuous multi-prompt engineering approach with medical knowledge injection. *Preprint*, arXiv:2401.12988.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press.

Rose Mintzer-Sweeney and Lisa Charlotte Muth. 2024. Data vis do's and don'ts. Accessed: 2024-10-02.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A systematic review of data-to-text nlg. *Preprint*, arXiv:2402.08496.

Ratish Puduppully, Yao Fu, and Mirella Lapata. 2022. Data-to-text generation with variational sequential planning. *Transactions of the Association for Computational Linguistics*, 10:697–715.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Tyler Richards. 2023. *Streamlit for Data Science: Create interactive data apps in Python*. Packt Publishing Ltd.

Kay Schröder, Wiebke Eberhardt, Poornima Belavadi, Batoul Ajdadilish, Nanette van Haften, Ed Overes, Taryn Brouns, and André Calero Valdez. 2023. Telling stories with data – a systematic review. *Preprint*, arXiv:2312.01164.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. The prompt report: A systematic survey of prompting techniques. *Preprint*, arXiv:2406.06608.

Hongbo Shao, Roberto Martinez-Maldonado, Vanessa Echeverria, Lixiang Yan, and Dragan Gasevic. 2024. Data storytelling in data visualisation: Does it enhance the efficiency and effectiveness of information retrieval and insights comprehension? CHI '24, New York, NY, USA. Association for Computing Machinery.

M.S. Smith. 2023. Google and microsoft race to unveil ai-based search. `https://spectrum.ieee.org/chatgpt`. Accessed: 2024-12-18.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting domain-specific knowledge into large language models: A comprehensive survey. *Preprint*, arXiv:2502.10708.

Ningjing Tang, Jiayin Zhi, Tzu-Sheng Kuo, Calla Kainaroi, Jeremy J Northup, Kenneth Holstein, Haiyi Zhu, Hoda Heidari, and Hong Shen. 2024. Ai failure cards: Understanding and supporting grassroots efforts to mitigate ai failures in homeless services. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 713–732.

Jingrong Tong and Landong Zuo. 2021. The inapplicability of objectivity: Understanding the work of data journalism. *Journalism Practice*, 15(2):153–169.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2024. Large language model benchmarks do not test reliability. In *Neurips Safe Generative AI Workshop 2024*.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2019. Learning neural templates for text generation. *Preprint*, arXiv:1808.10122.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2025. Number cookbook: Number understanding of language models and how to improve it. *Preprint*, arXiv:2411.03766.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659.

Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Zhenpeng Zhao and Niklas Elmqvist. 2023. The stories we tell about data: Surveying data-driven storytelling using visualization. *IEEE Computer Graphics and Applications*, 43(4):97–110.