

Can we demonstrate the centroid method is better than using an LLM directly to do similar coding tasks?)

Name Surname 1, Name Surname 2

June 26, 2024

Abstract

Input abstract here. The abstract should be short, and contain a summary of what is to be expected in the report. Some key findings is recommended to be highlighted in the abstract.

1 Introduction

- Some information about the purpose of the research
- Some information about Inter Rater Reliability and Cohen's Kappa test
- Discuss some information about reliability, cost,

An introduction is important to ensure the background of the report is sufficiently explored, and to give the readers up-to-date information before jumping into the content of the report. The main parts of an introduction should be a background of the content that will be covered in the report, the motivation of the project or purpose of the report, and finally the outline of the report. The outline of the report is important to let readers know what to expect by reading the report, what sections will be covered and a brief outlook on what it entails.

The page margins can be set under the package *geometry*. It is advisable to look up the margin requirements of your department.

If you are not familiar with concepts like styles, captioning, cross-referencing, and how to generate tables of contents, figures etc. in LaTeX, the Overleaf guides are a useful start at: https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes.

Otherwise, a lot of help can be obtained through Stack Exchange, where it is a forum that people get help from others regarding L^AT_EX writing.

Other useful tools are

- Detexify, which can convert your drawings to L^AT_EX commands,
- Mathpix tools, which you can use to screenshot an equation and it will convert into L^AT_EX equation form to be copied into your .tex document,
- Grammarly is a great help for grammar assistance on overleaf.

1.1 Sectioning

This is a subsection. You can also use sub-subsections, although it isn't recommended.

1.1.1 Subsubsection

Example of a sub-subsection. It isn't recommended to nest one more layer. However, if needed, the command is `\paragraph{insert section name}`. It should not be numbered.

Finally, make sure to outline the report at the end of the introduction section.

2 Background

This section should highlight the basis of the report, containing specific concepts used for the contents of the report. An alternative section title would be “**Theory**”. This should include any background understanding required for the content of the report, any literature review done and so on. Sometimes, this can be combined with the introduction, however, generally, it will look neater to have a section dedicated to the theory or background of the research or experimental analysis done.

This is different from the methodology which outlines the steps done to do the experimental work. Consider this to be an outline of the theoretical work done for the experimental (or simulation) work which will be outlined in the methodology section.

2.1 Writing tips

Generally, a scientific report should have justified text, as it looks neater, the right side of the report is not looking jagged. The font size should be 11pt to 13pt, but generally should be kept at 12pt. The recommended font type is either Arial or Calibri. Times New Roman is still accepted and generally the preferred font type, although recent talks about how it is less reader-friendly make it less appealing.

Given that Ireland uses UK English, please make sure any spellings should follow UK English spelling. If any information can be presented through paragraphing, it is recommended to do so. It is not recommended to use bullet or numbered points. It should be used sparingly, even though it might be easier to read.

Please also bear in mind that the main text should be within the margins. Any figures, equations and tables should still be within the margins. The following subsections will consider the specific points that should be taken note of.

2.2 Typography

Be mindful that the capitalisation of the first letter is only for given names. For example, although LED is a generally understood term, and the abbreviations are in capital letters, it should be written as light-emitting diode instead of capitalising the first letters. Any abbreviation should use uppercase letters (e.g. greenhouse gas (GHG)). If a few letters are used to represent an abbreviated word, lowercase letters can be used, such as Abbrev. for abbreviation or Pop. for population.

Typos could be mitigated with proofreading (especially on \LaTeX where the code editor and the PDF presented might have different outcomes). The use of Grammarly will help in checking grammatical and spelling errors if needed, especially for \LaTeX .

2.2.1 Scientific Writing Conventions

The **bold** style is not recommended to be used for emphasis. Excessive (or any) use of **bold** in text is frowned upon. Emphasis is usually done using *italics*. But again, please use it sparingly.

Italics in mathematical equations are also very particular. By convention, *italics* should only be used for mathematical terms (e.g. l for length, m for mass), while units should be in normal style (e.g. m for metre, and kg for kilogram). On \LaTeX , the default style in mathematical equations are *italics*, so if any units were to be introduced in the equation, please remember to remove the *italics*.

On that note as well, superscripts and subscripts should not be overlooked. Writing m^2 is different from m^2 , as well as V_{oc} from V_{OC} . When writing symbols, it is generally recommended to keep one symbol to represent one quantity, while using subscripts to specify the specific quantity it represents. For example, for the population of Dublin, *PopulationOfDublin*, *PopulationDublin* or *PopDublin* are not recommended. Instead, it is recommended to use P_{Dublin} or even N_D . Given that unknowns or quantities can be represented by a single symbol or letter, *Pop* could be an abbreviation for ‘population’ or three unknowns under implicit multiplications. This will be further discussed in subsection 2.3.

2.3 Equations and Numbers

Whenever an equation is introduced into the report, it must be useful to understand the mathematical work done in the report. Any mathematical equations used during the process of the research are recommended to be included in the report (although simple equations that are generally well-understood can be omitted).

A good practice is not to write the full name of the quantity in the equation but to introduce a term to represent it in the equation. This can be seen for example using Einstein's mass-energy equation in (1) and (2):

$$Energy = mass \times (Speed of Light)^2, \quad (1)$$

$$E = mc^2, \quad (2)$$

where E is energy, m is mass and c represents the speed of light.

Once terms are introduced as seen in (2), the equation is easier to read. Note also that any terms that have been introduced are fully described, and in a paragraph instead of bullet points. The equations are also treated as part of the sentence instead of a separate entity like figures or tables. That is why a punctuation mark was used at the end of each equation.

The equations should also be labelled so that it is easily referred to when it is used. Thus, instead of inputting values in the equations, it is recommended to leave them out and conclude the calculations in text. For example, the gradient of a linear function, m can be written as

$$m = \frac{y - c}{x}, \quad (3)$$

where x and y are independent variables with a linear correlation and c is the y-axis intercept. For a directly proportional relationship (i.e. $c = 0$), and a coordinate of (2.0, 7.0) was found, by using (3), m is calculated to be 3.5.

Although it is fine to include the units in the equation, it is generally better to put it in the description of the terms. For example, energy, E in Wh can be calculated as

$$E = P \times t, \quad (4)$$

where P is the power in watts (W) and t is the time taken in hours (h). Bonkers.ie stated that the average household energy usage in a year is 4200kWh a year, while 50% of the households use between 3150–5250kWh a year [1]. Note that even if a range is written, the units should only be at the end, and the dash should be an en dash.

Generally, it is recommended to use indices when writing values. It is easier to understand 3.1×10^{-3} compared to 0.0031. Likewise, since the purpose is for ease of reading and interpreting, if any indices are used, the indices should be consistent throughout the variable for easy comparison.

2.4 Citations and Reference

When citing a document, the reference should contain as much information as possible. The reference should be consistent as well, with the order of information and style of writing. Some report has an inconsistent style of writing the authors. If initials are used for first names, all the authors should only have initials for first names. Most article database (or even journal websites) will have the option to download the citation as BibTeX, which then can be copied into the bibliography .bib file. The reference should be at the end of the document, before the appendix.

When citing in the main text, the common two ways of doing it is by numbers (generally Vancouver style [2]) or by author-year (generally Harvard style [3]). When mentioning the name of the author during the citation, *et al.* should be used and in italics. If using numbered styles, the number can appear after the name of the author or at the end of the sentence.

One source that is frowned upon as a reference is Wikipedia. Given the nature of Wikipedia being open for editing by the public, the generally quick-changing nature of the Wikipedia page makes it not a suitable source of reference. However, Wikipedia is still useful to obtain resources from their citation list. Although Wikipedia should not be referenced, Wikimedia figures can be referenced, although not recommended as well.

2.4.1 Plagiarism

Changing a few words in a sentence from a resource is insufficient for paraphrasing. Paraphrasing requires a rewrite of the content in your own words, not a slight modification of a given work. Take note as well, although ChatGPT can be a good tool to help paraphrase a sentence, make sure that the content has not changed and produce a conceptually false sentence. **Please look up your department policies regarding the use of AI writing tools in your report writing.**

At the same time, picking up a full sentence (or a huge part of a sentence) from your sources are also considered plagiarism, even though the sentence is short and not the full content of the report. Although not recommended, quoting should be used if a full quote is picked up to be used in the report. A general guide is that a single quotation mark is for a direct quote from the source, while 'double quotation marks are usually used for quoting direct speech' [2, p. 6]. Also, take note of the punctuation used as the open quotation mark is a different symbol to the close quotation mark in L^AT_EX.

If paraphrasing can be done, that would be much more recommended than a direct quote. However, if the message cannot be transmitted through paraphrasing, a limited amount of direct quoting can be done.

Finally, no matter where you have obtained your data, please cite it. Even if the source is the lecture content, please cite the source. As long as it is not from you, you have to cite it. If you looked up more information to analyse your data, *cite where you obtained the information that led you to the conclusion you have made*. In this lab report, the most number of reference per report is one, which highly unlikely that all the information in the report is your own output without any external support or reference.

One final note, that **self-plagiarism is still plagiarism**, and you should never submit any work that you have submitted before. If it showed to be useful for the current report, it has to be cited and paraphrased as much as possible.

3 Methodology

Elaborate in detail the methods you used to collect the data. This should also include any potential uncertainty and errors. Any relevant equations should be explained either here or in the background section.

Typically, scientific writing should be in the passive voice. The use of active voice has been more and more accepted in scientific writing, however, it is generally better to be in a passive voice. A passive voice removes the author from the work, which makes the work done feel more generalised, and should be replicable by anybody. Since scientific work should be factual by nature, the results should be the same regardless of the individual doing the experiment. Although this is not always true, especially considering bias, it is generally recommended to leave the subject out of the report.

3.1 Constructed Dataset

The process started with the creation of a dataset centered around the question "Why did you choose the Honours program". 25 fictitious responses to this question were made, which made up the dataset. The dataset was then coded into one of the four categories "Career Opportunities", "Social Life", "Learning" and "Other", which was used for the responses that did not fit any of the aforementioned categories. This coding formed what we will henceforth refer to as the ground truth.

Based on this dataset, ChatGPT UiO was prompted in different ways.

- (In the beginning), ChatGPT UiO was simply asked to sort the responses into categories using the following prompt: (prompt).
- The prompt was then updated to the following: (prompt)
- and (prompt)

Finally ChatGPT UiO was asked to sort the data into the four (predefined) categories using the following prompt: (prompt)

This process was repeated three times. The data was collected, sorted into a list and each category was assigned a number for ease of computation.

To measure the Inter Rater Reliability, Cohen's Kappa Score was used using the built-in functionality in the python package *scikitlearn.metrics*.
(reference to the code used)

We then used the centroid method on this dataset using the code provided from the Colab. - have to specify the parameters used

3.2 News Headlines Dataset

For this project we used a pre-categorized dataset from Hugging Face <https://huggingface.co/datasets/okite97/news-data>, containing news headlines which had been labeled with the following categories: "politics", "business", "health", "sports", "tech" and "entertainment". The dataset had already been split into a training- and test-dataset, and we used the test-dataset which consisted of 828 headlines and used the first 500.

We then proceeded to manually prompt ChatGPT UiO by asking it to categorize the first 500 headlines by giving it chunks of 50 at a time and starting a new chat per chunk. This prompting was done using the following prompt: (prompt)

The prompt had to be somewhat modified in the process, since the format of the output from GPT varied, despite the specifications in the prompt. (Explain further)

The dataset of the 500 first datapoints were coded by ChatGPT UiO three separate times and each run was collected into a single document. This text was processed using the following code:

(code)

The produced lists were then used as a basis for calculating the Inter Rater Reliability using Cohen's Kappa test using the same procedure as we did in (first project).

We also used the centroid method to obtain a categorization. The centroid method gives its result as a distribution between the different categories. We wrote a script to obtain only the most probable category two different ways. First we set $\alpha = 150$ to shift the weight to get a value greater than 0.50 for all of them and compared this categorization to the ground truth using Cohen's Kappa Score using the following code:

(code)

We then set $\alpha = 15$ and excluded the data where the greatest value in the distribution was smaller than 0.50, and constructed a ground truth where we excluded the same values, and used Cohen's Kappa Score to compare the two using the following code:

(code)

Llama

Categorization was done using Llama3 but using prompting in the terminal using a Powershell script, and using 3"cloud server".

The prompting in the Powershell was done using the following code:

(code) The prompting with the 3"cloud server" was done using the following code:

(code)

3.3 Third "project"

4 Results

4.1 Constructed Dataset

When asked to simply sort the data into categories ChatGPT UiO sorted every response into its own category

Run	Ground truth	Run 1	Run 2	Run 3
Ground truth	1	0.7635933806146572	0.8842592592592593	0.8842592592592593
Run 1	0.7635933806146572	1	0.7530864197530864	0.7530864197530864
Run 2	0.8842592592592593	0.7530864197530864	1	0.8798076923076923
Run 3	0.8842592592592593	0.7530864197530864	0.8798076923076923	1

Table 1: Placeholder

4.2 News Headlines Dataset

ChatGPT UiO:

Run	Ground truth	Run 1	Run 2	Run 3
Ground truth	1	0.8968636487125565	0.7811913027230619	0.8942869655828564
Run 1	0.8968636487125565	1	0.8183344707886739	0.9497358619545713
Run 2	0.7811913027230619	0.8183344707886739	1	0.8231740262291861
Run 3	0.8942869655828564	0.9497358619545713	0.8231740262291861	1

Table 2: Cohen's Kappa Score for three different runs of categorization done by ChatGPT UiO.

Llama3:
Centroid Method:

5 Discussion

Sources of Error

1. had to change the prompt somewhat during the prompting of ChatGPT

Run	Ground truth	Run 1	Run 2	Run 3
Ground truth	1	0.8748354085622594	0.869899822863605	0.8650661176023748
Run 1	0.8748354085622594	1	0.8183344707886739	0.9525706697021438
Run 2	0.869899822863605	0.9550195664885774	1	0.9650587013816788
Run 3	0.8650661176023748	0.95257066970214383	0.9650587013816788	1

Table 3: Initial capital structures of large projects (\$1bn.+) (*Finnerty, 2013*)

Run	Ground truth
Centroid Method	0.775840597758406
Excluded Centroid Method	0.8212882389320273

Table 4: Initial capital structures of large projects (\$1bn.+) (*Finnerty, 2013*)

2. ChatGPT sometimes gives its response in different formats
3. Llama3 at some point gave the answer as something other than just the category
4. first dataset - too small for the centroid Method

Placeholder

Discussion of findings, interpretation of the results

- The centroid method seemed to miss quite a lot in the first set, probably because the set was too small

Discussion of reliability, transparency, cost...

6 Conclusion

A clear and concise conclusion, containing key takeaways of the article or report. It is useful to give the readers a huge summary of what they just read, highlighting the main takeaway from the report. It should provide the key information that the author wishes the reader to remember most when reading the report. Although this sounds very similar to the abstract, this is generally true, but make sure the content of both the abstract and the conclusion do not overlap. Basically, the main body should have an elaborate discussion while summarised in the conclusion.

References

- [1] Daragh Cassidy. What is the national average energy consumption?, n.d. URL <https://www.bonkers.ie/guides/gas-electricity/national-average-energy-consumption/>. [Accessed 19 November 2023].
- [2] Library Services: Imperial College London. Citing and Referencing: Vancouver Style, n.d.. URL <https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/library/public/IMPP10650-College-Vancouver-Guide-230822-WEB.pdf>. [Accessed 19 November 2023].
- [3] Library Services: Imperial College London. Citing and Referencing: Harvard Style, n.d.. URL <https://www.imperial.ac.uk/media/imperial-college/administration-and-support-services/library/public/IMPP10649-Harvard-Guide-230822-WEB.pdf>. [Accessed 19 November 2023].

A First Appendix

This is the first appendix. For more appendices, add \section{*appendix title*} for another appendix.