What I’ve learnt about writing OII papers…

Ten marked papers

Term 1

- Foundations
- Data Science / ML
- Statistics

Term 2

- Research Design 1 (?)
- Research Design 2 (?)
- Research Design 3 (?)
- Frontiers
- NLP
- Applied ML

Term 3

- Thesis

Disclaimer: These are all my own views and not necessarily those of the people marking your work. A lot of these topics are subjective. These comments are all for statistics and not other papers. All examples are from my own papers and feedback is real feedback I received.

Plan

1. Structuring your paper
2. Figures and tables
3. Writing style
4. General tips
5. Model essays
6. Questions

- My first paper was my worst and my last was my best
- Almost linearly got better each paper
- This presentation shows my ‘quick tips’ that I can give you

1

Structuring Your Paper

1 Use the traditional scientific structure

Title Page: This should briefly but explicitly describe the purpose of the report.

Summary (Abstract): Often only 100 to 300 words, the abstract generally provides a broad overview and is never more than a page. It describes the essence, the main theme of the paper. It includes the research question posed, its significance, the methodology, and the main results or findings. Footnotes or cited works are never listed in an abstract. Remember to take great care in composing the abstract. It's the first part of the paper the instructor reads.

Introduction: The introduction sets the scene for the main body of the report. The aims and objectives of the report should be explained in detail.

1 Literature review/background section?

"Perhaps due to the length of this work, there was no need for separate sections on Introduction and Background."
- AAS Feedback

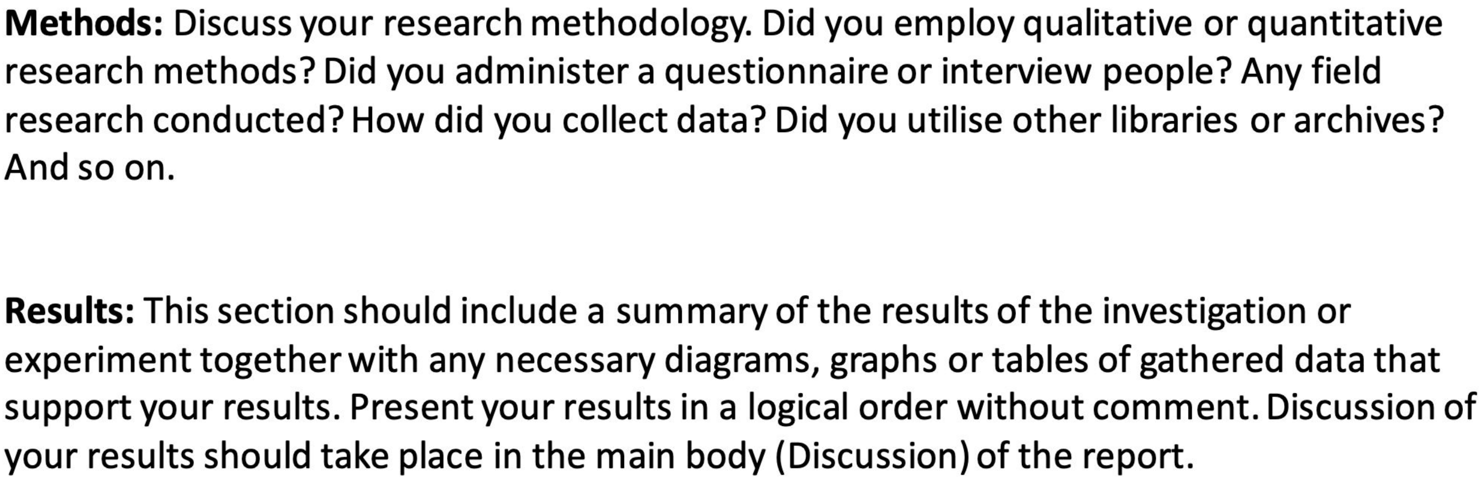
But in other papers potentially…

1 Literature review/background section?

Generally called “Related work” in most ML papers. New(ish) trend to put the literature review around the discussion or in the Appendix.

Would **strongly advise against** doing this for your papers.

1

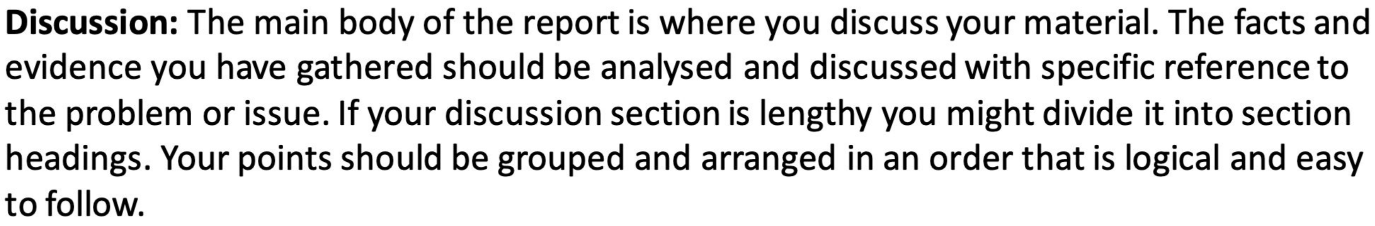


Methods: Discuss your research methodology. Did you employ qualitative or quantitative research methods? Did you administer a questionnaire or interview people? Any field research conducted? How did you collect data? Did you utilise other libraries or archives? And so on.

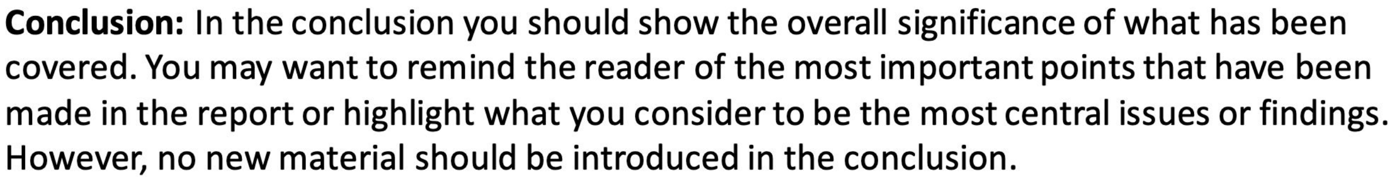
Results: This section should include a summary of the results of the investigation or experiment together with any necessary diagrams, graphs or tables of gathered data that support your results. Present your results in a logical order without comment. Discussion of your results should take place in the main body (Discussion) of the report.

You should not discuss your results them in the context of the question here. Save that for the discussion. A good rule of thumb on this is that you shouldn't be discussing literature.

1


Discussion: The main body of the report is where you discuss your material. The facts and evidence you have gathered should be analysed and discussed with specific reference to the problem or issue. If your discussion section is lengthy you might divide it into section headings. Your points should be grouped and arranged in an order that is logical and easy to follow.

Link back to the literature. Do not introduce any new results. All results should be in the results section. See example on the next page. Avoid speculation!

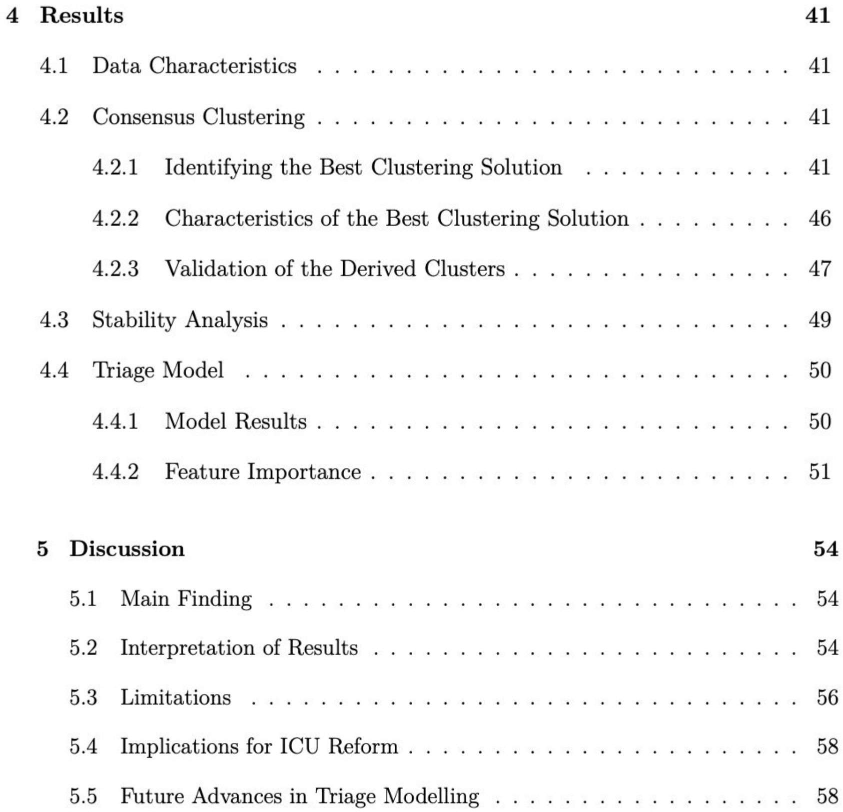

Conclusion: In the conclusion you should show the overall significance of what has been covered. You may want to remind the reader of the most important points that have been made in the report or highlight what you consider to be the most central issues or findings. However, no new material should be introduced in the conclusion.

Do not introduce any new ideas.

1

Structure of my thesis

OXFORD INTERNET INSTITUTE



- I had 3 separate parts of my methodology/results
- Discuss each in turn in the results
- Discuss each in turn in the discussion.

- Much better than going results for model 1, discussion for model 1, results for model 2, discussion for model 2… even though that is how I processed it in my head when working on the paper

1 Structuring the discussion

1. Main point. Relate to existing research. Lots of literature.

2. Potentially further sections with more detail

3. Limitations! Very important to include. It won't undermine your work.

"…although it would have been interesting to see this argument further developed in the limitation section."
- Applied ML

"The discussion section contextualises the findings and implications, and it openly discusses the limitations of the study, which can inform future research."
- Thesis

1

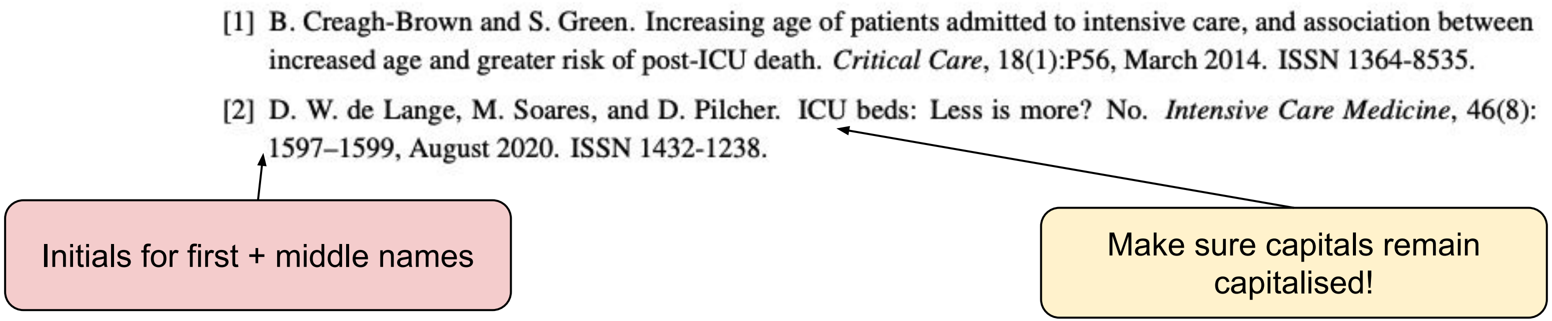
Appendices: An appendix contains material that is appropriate for enlarging the reader's understanding, but that does not fit very well into the main body of the paper. Such material might include tables, code, pseudocode, charts, summaries, questionnaires, interview questions, lengthy statistics, maps, pictures, photographs, lists of terms, glossaries, survey instruments, letters, copies of historical documents, and many other types of supplementary material. A paper may have several appendices.

Bibliography: Your bibliography should list all published sources referred to in your report. There are different styles of using references and bibliographies. Each reference must be used in the report.

1 Bibliography

"…clear errors in your bibliography"

- Surprisingly obvious if your bibliography is formatted incorrectly (the examiner will check!)
- Follow the OII's guidance on structuring bibliographies, but I normally do…



[1] B. Creagh-Brown and S. Green. Increasing age of patients admitted to intensive care, and association between increased age and greater risk of post-ICU death. *Critical Care*, 18(1):P56, March 2014. ISSN 1364-8535.

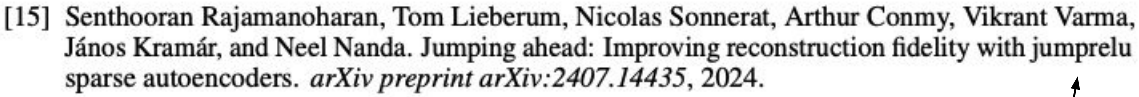
[2] D. W. de Lange, M. Soares, and D. Pilcher. ICU beds: Less is more? No. *Intensive Care Medicine*, 46(8): 1597–1599, August 2020. ISSN 1432-1238.

Initials for first + middle names

Make sure capitals remain capitalised!

1 Bibliography

A bad example…



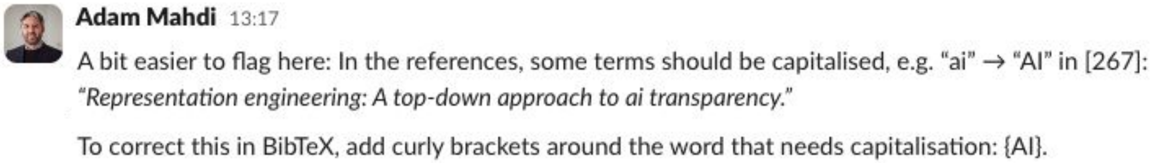
[15] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Cited the arXiv version but this paper is at NeurIPS

Maybe should have used initials?

Should be JumpReLU

1 Bibliography



Adam Mahdi 13:17

A bit easier to flag here: In the references, some terms should be capitalised, e.g. “ai” → “AI” in [267]: *“Representation engineering: A top-down approach to ai transparency.”*

To correct this in BibTeX, add curly brackets around the word that needs capitalisation: {AI}.

1 Bibliography

“…clear errors in your bibliography”

- Use a **reference manager** and then can create bibtex citations easily. E.g. Zotero.
- Use **LaTeX package** to help with consistency. Biblatex or natbib
- LaTeX automatically adds, edits and formats the relevant citations

1 Bibtex settings I used in my thesis

```
% Bibtex settings I used in my thesis
\usepackage[style=apa, sorting=nyt, backend=biber, maxcitenames=2, useprefix, doi=false, isbn=false,
giveninits=true, uniquename=false]{biblatex}

% Changes the title to References
\newcommand*{\bibtitle}{References}

% Load the bib file
\addbibresource{Thesis.bib} % Note that \addbibresource is recommended over \bibliography in biblatex

% Prevents the 'note' section being printed
\AtEveryBibitem{\clearfield{note}}

\AtEveryBibitem{
 \clearfield{annote}
}
```

1 What is BibTeX?

- Reference management system. Supported by softwares like Zotero and Mendeley
- Can also generate straight from Google Scholar
- Check Zotero / Google Scholar output! Often Zotero is slightly wrong

```
@article{vaswani2017attention,
  title={Attention is all you need},
  author={Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez,
Aidan N and Kaiser, {\L}ukasz and Polosukhin, Illia},
  journal={Advances in neural information processing systems},
  volume={30},
  year={2017}
}
```

1

I would suggest trying to stick to the traditional scientific structure as closely as possible

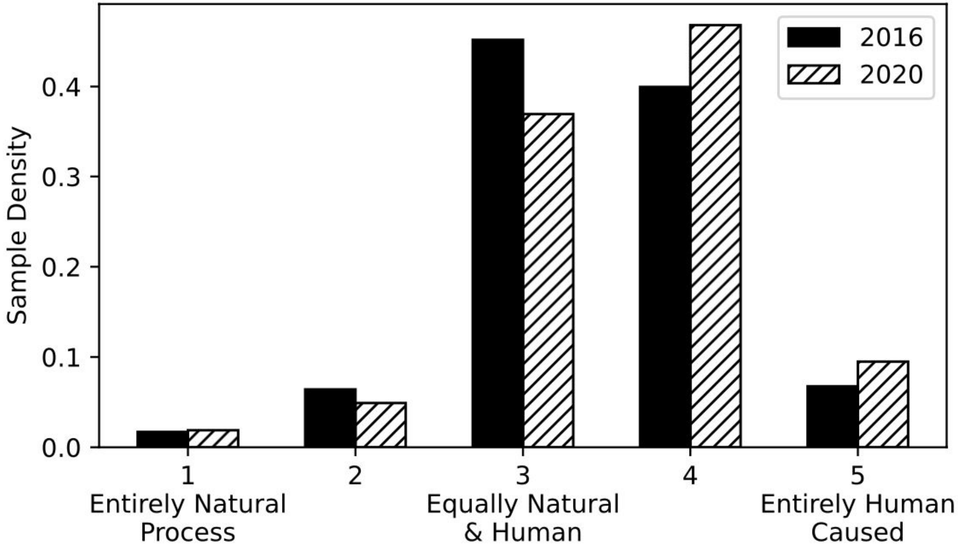
2 Figures and Tables

2 Tips

1. Use (minimal) colour for figures
2. Use consistent colour throughout your paper
3. Title and caption underneath figures
4. Make captions **self-sufficient**
5. Use high-quality images
6. LaTeX table generator? Some people swear by this. ChatGPT also good.

2

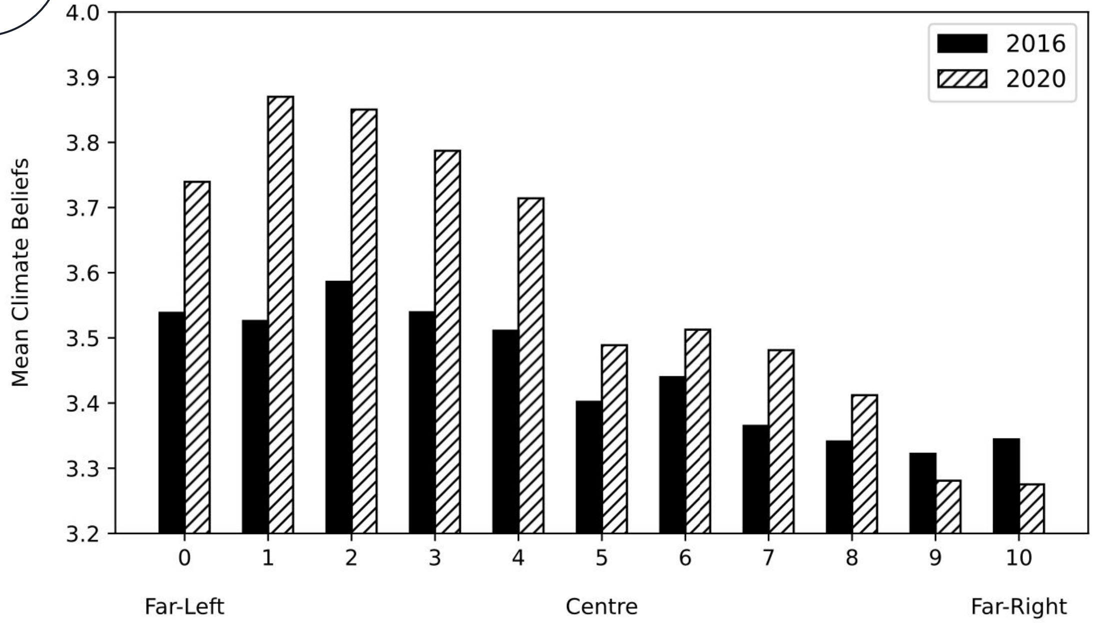
Figure 1: Beliefs About the Causes of Climate Change in 2016 and 2020.



Pros and Cons…

2

Figure 2: Climate Beliefs Across the Political Spectrum in 2016 and 2020



Pros and Cons…

❌ Caption not self-contained
❌ Shading not ideal
❌ Title capitalisation wrong
❌ TItle bolding wrong

“the presentation would benefit if the tables and figures had more information in the caption making them more self-contained” - AAS Feedback

2

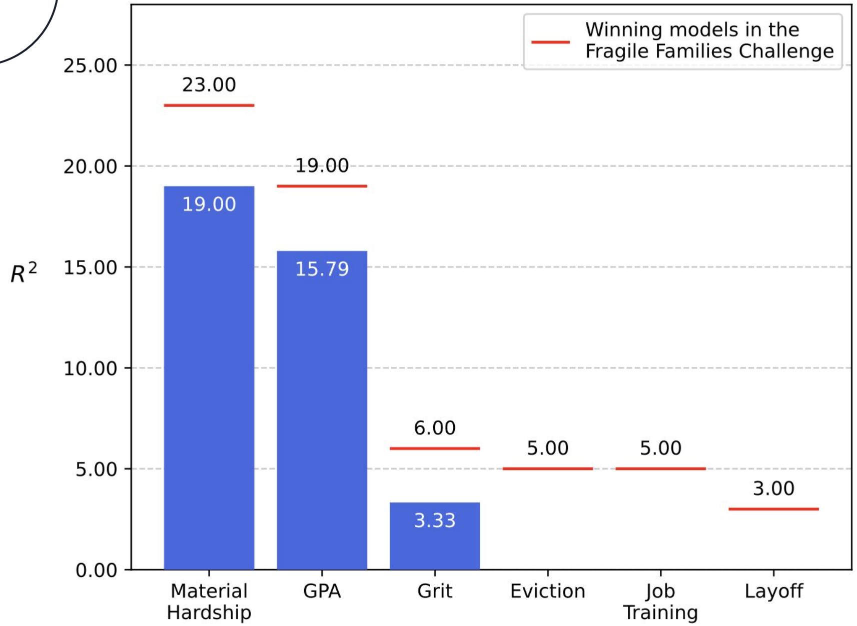


Figure 2: R^2 scores for the best models with scores from the winning models in the original challenge shown in red. Scores for eviction, job training and layoff did not beat the mean baseline and are not shown.

Pros and Cons…

✅ Nice colour
✅ Nice design
✅ Use of bold in title okay

❌ No title…?

“Figure 2 and 3 are very well designed” - NLP Feedback

2

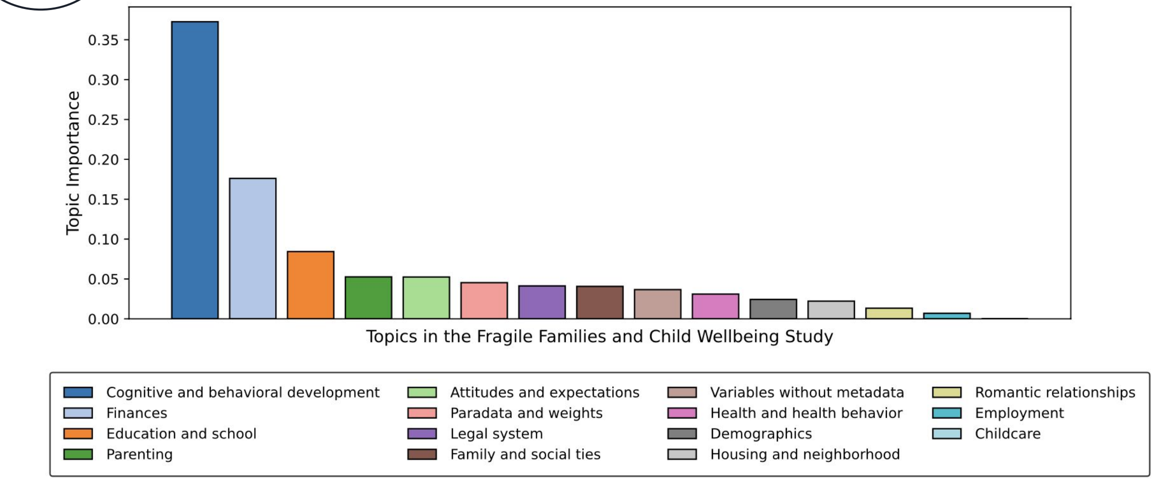


Figure 3: Topic importance for predicting GPA in children aged 15. The topics aggregate features in the Fragile Families and Child Wellbeing Study. Topic importance is a metric combining feature importance from LASSO and GBM.

Pros and Cons…

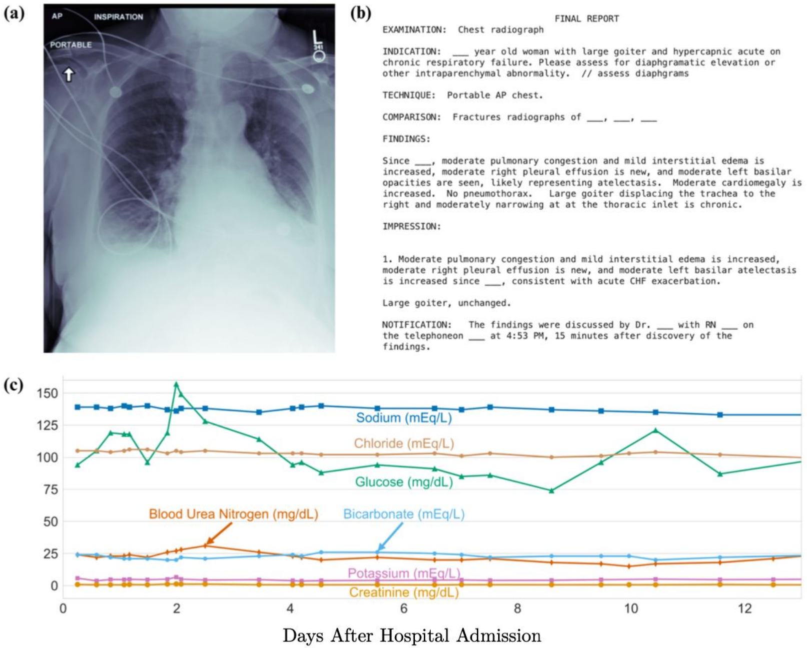
✅ Use of bold in title okay

❌ Colours… bizarre
❌ Weird formatting
❌ No title
❌ Capitalisation of axis wrong

"The figures are good (though in publishing any future work, the student should make sure to use consistent colour coding across figures)" - NLP Feedback

2

Pros and Cons…



```
                                 FINAL REPORT
EXAMINATION:  Chest radiograph

INDICATION:  ___ year old woman with large goiter and hypercapnic acute on
chronic respiratory failure. Please assess for diaphgramatic elevation or
other intraparenchymal abnormality.  // assess diaphgrams

TECHNIQUE:  Portable AP chest.

COMPARISON:  Fractures radiographs of ___, ___, ___

FINDINGS:

Since ___, moderate pulmonary congestion and mild interstitial edema is
increased, moderate right pleural effusion is new, and moderate left basilar
opacities are seen, likely representing atelectasis.  Moderate cardiomegaly is
increased.  No pneumothorax.   Large goiter displacing the trachea to the
right and moderately narrowing at at the thoracic inlet is chronic.

IMPRESSION:


1. Moderate pulmonary congestion and mild interstitial edema is increased,
moderate right pleural effusion is new, and moderate left basilar atelectasis
is increased since ___, consistent with acute CHF exacerbation.

Large goiter, unchanged.

NOTIFICATION:   The findings were discussed by Dr. ___ with RN ___ on
the telephoneon ___ at 4:53 PM, 15 minutes after discovery of the
findings.
```

Figure 5.1: Unused Modalities in the Triage Model. (a) An example of a chest X-ray for a hypercapnic patient (Johnson et al., 2018; Johnson et al., 2022) **(b)** The free-text radiology report accompanying the X-ray **(c)** An example of laboratory readings taken across a single patient's hospitalisation (Johnson et al., 2023). Currently, the triage model only considers the reading taken closest to ICU admission and incorporating time series data would enhance model performance.

✅ Figure title
✅ Good use of bold
✅ (a), (b) and (c) labelled

❌ Capitalisation wrong
❌ Kind of confusing…

2

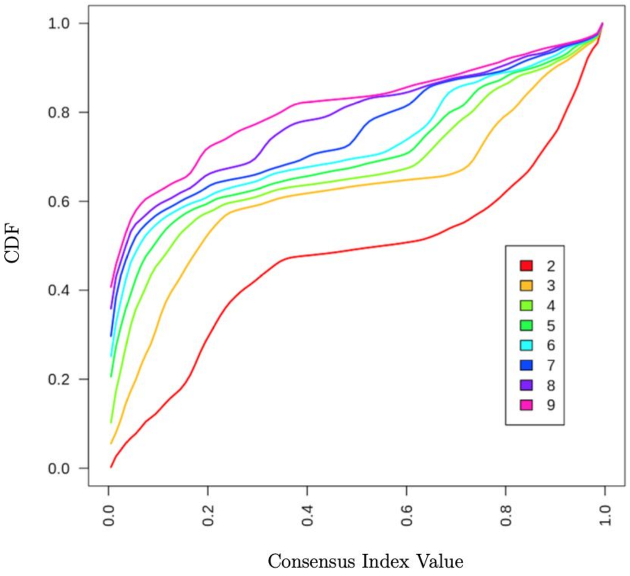
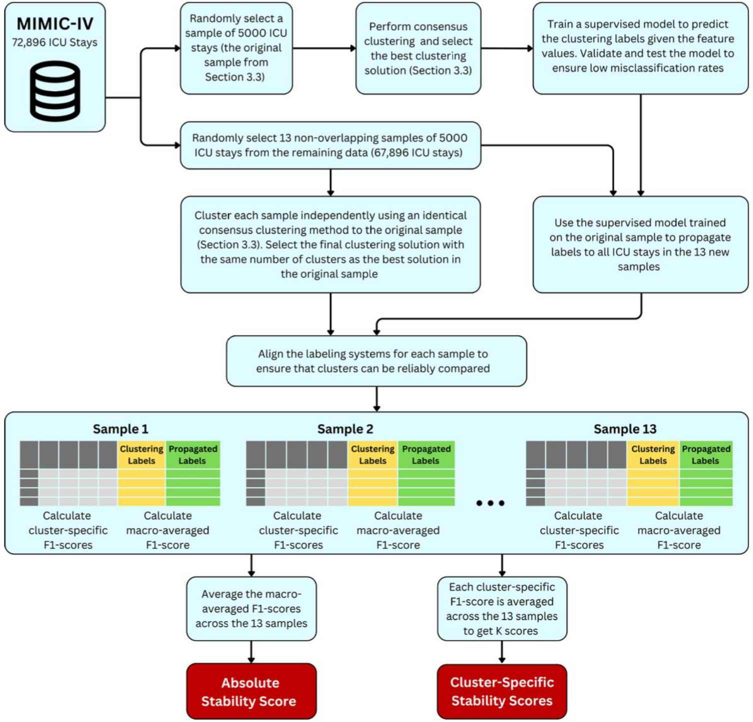


Figure 4.2: Cumulative Distribution Functions for the Clustering Results. A step-like function with more weight near 0.0 and 1.0 is indicative of a more robust clustering. Solutions with a lot of weight in central portions of the consensus index range indicate clusters with unstable membership. This figure suggests $K = 3$ is the most stable solution, although $K = 2$ and $K = 4$ are also similar.

“Weird use of colours…”

2



MIMIC-IV
72,896 ICU Stays

Randomly select a sample of 5000 ICU stays (the original sample from Section 3.3)

Perform consensus clustering and select the best clustering solution (Section 3.3)

Train a supervised model to predict the clustering labels given the feature values. Validate and test the model to ensure low misclassification rates

Randomly select 13 non-overlapping samples of 5000 ICU stays from the remaining data (67,896 ICU stays)

Cluster each sample independently using an identical consensus clustering method to the original sample (Section 3.3). Select the final clustering solution with the same number of clusters as the best solution in the original sample

Use the supervised model trained on the original sample to propagate labels to all ICU stays in the 13 new samples

Align the labeling systems for each sample to ensure that clusters can be reliably compared

Calculate cluster-specific F1-scores

Calculate macro-averaged F1-score

Average the macro-averaged F1-scores across the 13 samples

Each cluster-specific F1-score is averaged across the 13 samples to get K scores

Absolute Stability Score

Cluster-Specific Stability Scores

Figure 3.5: Schematic Diagram of the Stability Analysis Method. This method generates two stability metrics which are highlighted in red: The absolute stability score, which is the average macro-averaged F1-score across the 13 samples, and the cluster-specific stability scores, which are cluster-specific F1-scores averaged over the 13 samples.

“Why blue? Why red? Why so many arrows?”

- Make sure it is consistent with what most academic papers do

2 Infographics

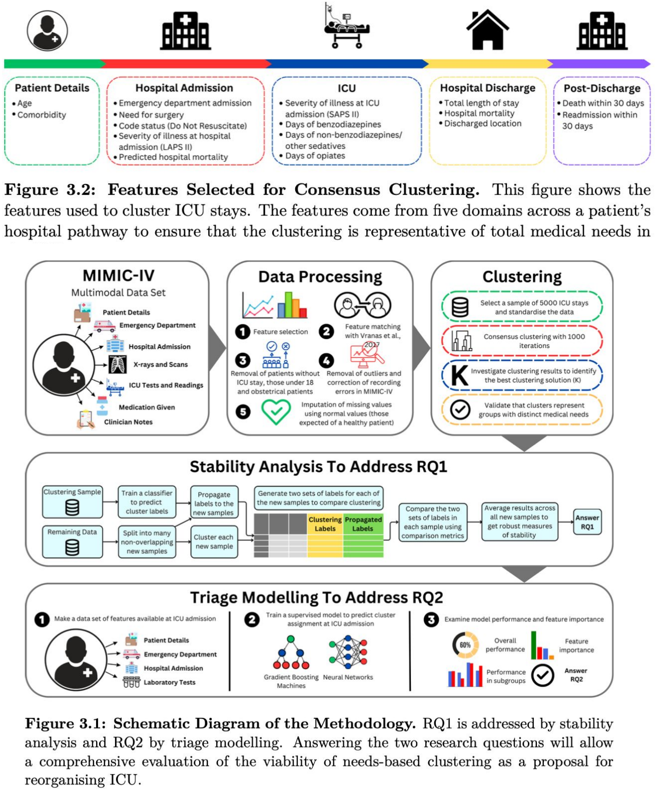

Patient Details
• Age
• Comorbidity
Hospital Admission
• Emergency department admission
• Need for surgery
• Code status (Do Not Resuscitate)
• Severity of illness at hospital admission (LAPS II)
• Predicted hospital mortality
ICU
• Severity of illness at ICU admission (SAPS II)
• Days of benzodiazepines
• Days of non-benzodiazepines/ other sedatives
• Days of opiates
Hospital Discharge
• Total length of stay
• Hospital mortality
• Discharged location
Post-Discharge
• Death within 30 days
• Readmission within 30 days

Figure 3.2: Features Selected for Consensus Clustering. This figure shows the features used to cluster ICU stays. The features come from five domains across a patient's hospital pathway to ensure that the clustering is representative of total medical needs in

MIMIC-IV
Multimodal Data Set
Patient Details
Emergency Department
Hospital Admission
X-rays and Scans
ICU Tests and Readings
Medication Given
Clinician Notes
Data Processing
1 Feature selection
2 Feature matching with Vranas et al., 2017
3 Removal of patients without ICU stay, those under 18 and obstetrical patients
4 Removal of outliers and correction of recording errors in MIMIC-IV
5 Imputation of missing values using normal values (those expected of a healthy patient)
Clustering
Select a sample of 5000 ICU stays and standardise the data
Consensus clustering with 1000 iterations
K
Investigate clustering results to identify the best clustering solution (K)
Validate that clusters represent groups with distinct medical needs
Stability Analysis To Address RQ1
Clustering Sample
Remaining Data
Train a classifier to predict cluster labels
Split into many non-overlapping new samples
Propagate labels to the new samples
Cluster each new sample
Generate two sets of labels for each of the new samples to compare clustering
Clustering Labels
Propagated Labels
Compare the two sets of labels in each sample using comparison metrics
Average results across all new samples to get robust measures of stability
Answer RQ1
Triage Modelling To Address RQ2
1 Make a data set of features available at ICU admission
Patient Details
Emergency Department
Hospital Admission
Laboratory Tests
2 Train a supervised model to predict cluster assignment at ICU admission
Gradient Boosting Machines
Neural Networks
3 Examine model performance and feature importance
60%
Overall performance
Feature importance
Performance in subgroups
Answer RQ2

Figure 3.1: Schematic Diagram of the Methodology. RQ1 is addressed by stability analysis and RQ2 by triage modelling. Answering the two research questions will allow a comprehensive evaluation of the viability of needs-based clustering as a proposal for reorganising ICU.

"Quite a lot of boxes but I like it"

- Can be effective
- Use PowerPoint or Canva to make

2 Infographics

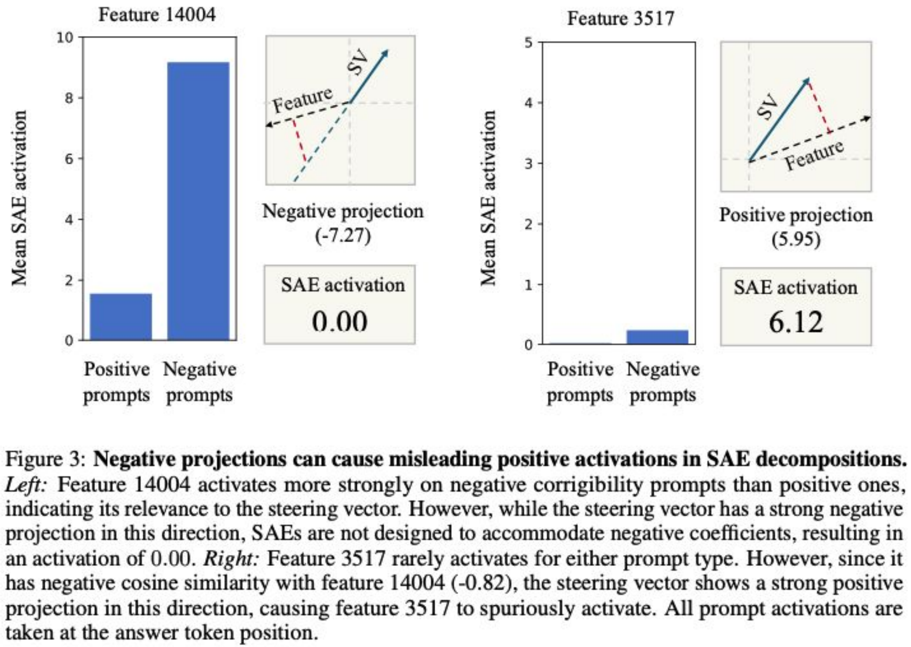
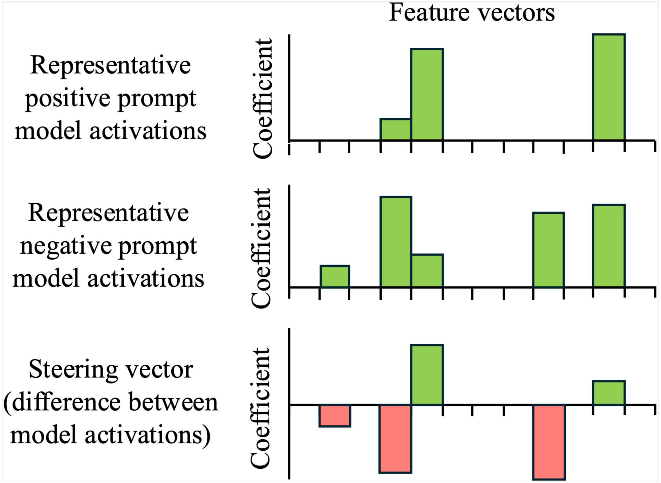


Figure 3: **Negative projections can cause misleading positive activations in SAE decompositions.** *Left:* Feature 14004 activates more strongly on negative corrigibility prompts than positive ones, indicating its relevance to the steering vector. However, while the steering vector has a strong negative projection in this direction, SAEs are not designed to accommodate negative coefficients, resulting in an activation of 0.00. *Right:* Feature 3517 rarely activates for either prompt type. However, since it has negative cosine similarity with feature 14004 (-0.82), the steering vector shows a strong positive projection in this direction, causing feature 3517 to spuriously activate. All prompt activations are taken at the answer token position.

✅ Nice title
✅ Self-contained

❌ Repeated y-axis label could probably be removed…

2 Avoid visualisation without real data



✅ Looks okay but…

❌ Bit cheap
❌ Not real data so not very convincing

3

Writing Style

3 Difference between an journal paper and an OII summative

Journal Paper

- Readers assumed to know the topic so can assume away the basics
- Aim in to show results/discussion so might be light on methods
- Might have journal-specific structure

OII Paper

- Readers have limited knowledge of topic so must prove you know the basics
- Aim is to show you understand the methods so heavier on methods
- Best to stick to the traditional structure

Different aims, different structure

3

Writing Tips

1. Write, write, write. Write without hesitation first, then edit.

2. Start early and write to help you think through problems

3. Write structures of arguments and sections on paper to get ideas out

3

Writing Style: Things not to forget when editing

1. Tenses! Look up how papers in your discipline are normally written. Really obvious if you mess this up

2. Make it clear and concise. Don’t be verbose.

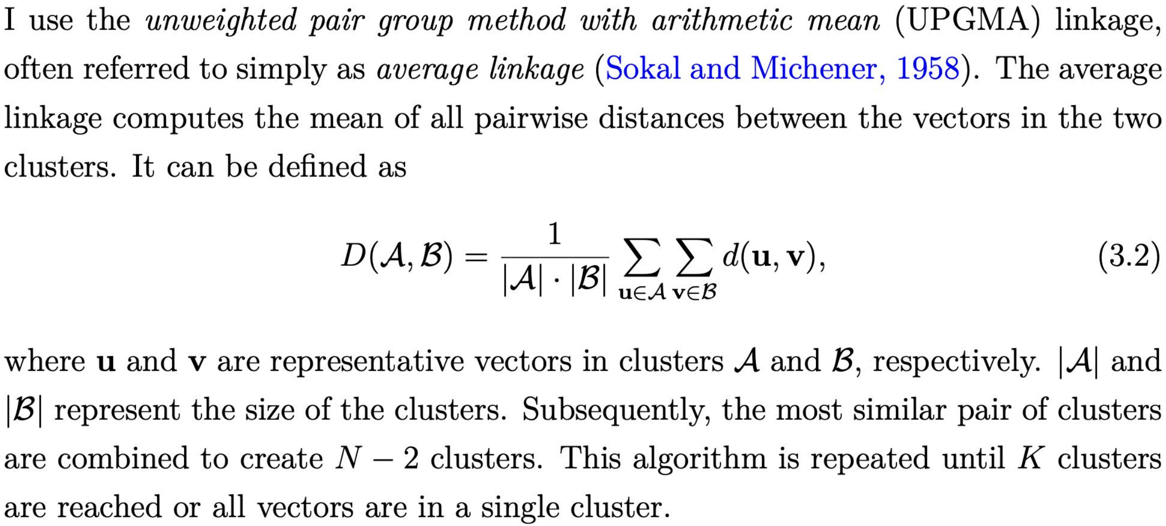
3. Consistent capitalisation/text size/formatting throughout the paper

3

Writing Style: Style Guides

1. *The Elements of Style*, Struck and White
2. *How to Write in Plain English*
3. *Politics and the English Language*, Orwell
4. *The Economist Style Guide*, good for word choice
5. *On Writing*, King
6. *The Sense of Style*, Pinker

3 Writing Style: Equations



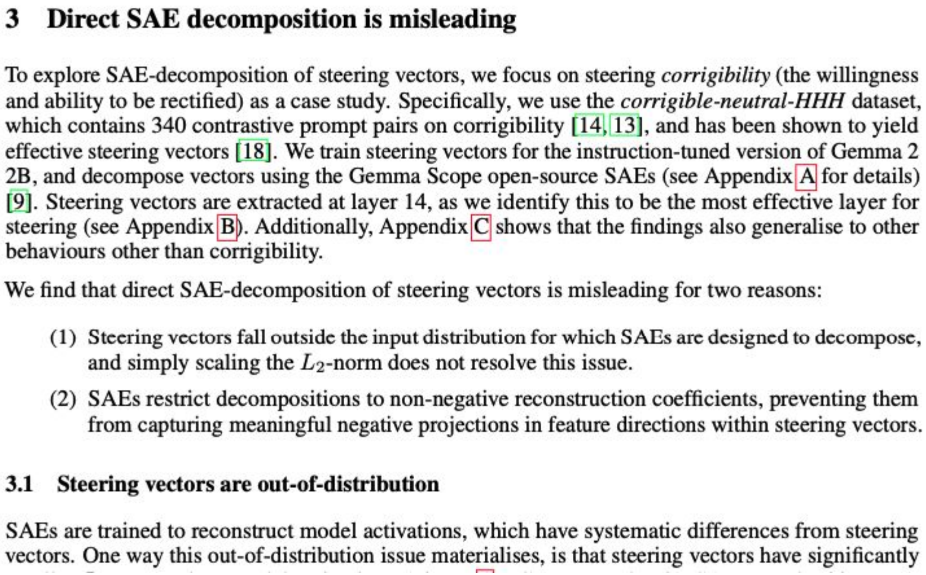
I use the *unweighted pair group method with arithmetic mean* (UPGMA) linkage, often referred to simply as *average linkage* (Sokal and Michener, 1958). The average linkage computes the mean of all pairwise distances between the vectors in the two clusters. It can be defined as

$$D(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{\mathbf{u}\in\mathcal{A}} \sum_{\mathbf{v}\in\mathcal{B}} d(\mathbf{u},\mathbf{v}), \tag{3.2}$$

where $\mathbf{u}$ and $\mathbf{v}$ are representative vectors in clusters $\mathcal{A}$ and $\mathcal{B}$, respectively. $|\mathcal{A}|$ and $|\mathcal{B}|$ represent the size of the clusters. Subsequently, the most similar pair of clusters are combined to create $N-2$ clusters. This algorithm is repeated until K clusters are reached or all vectors are in a single cluster.

- Often defined within the text
- No sentence breaks
- For big equation/proofs you can do stand alone equations

3 Writing Style: Capitals



3 Direct SAE decomposition is misleading

To explore SAE-decomposition of steering vectors, we focus on steering *corrigibility* (the willingness and ability to be rectified) as a case study. Specifically, we use the *corrigible-neutral-HHH* dataset, which contains 340 contrastive prompt pairs on corrigibility [14, 13], and has been shown to yield effective steering vectors [18]. We train steering vectors for the instruction-tuned version of Gemma 2 2B, and decompose vectors using the Gemma Scope open-source SAEs (see Appendix A for details) [9]. Steering vectors are extracted at layer 14, as we identify this to be the most effective layer for steering (see Appendix B). Additionally, Appendix C shows that the findings also generalise to other behaviours other than corrigibility.

We find that direct SAE-decomposition of steering vectors is misleading for two reasons:

(1) Steering vectors fall outside the input distribution for which SAEs are designed to decompose, and simply scaling the L_2-norm does not resolve this issue.

(2) SAEs restrict decompositions to non-negative reconstruction coefficients, preventing them from capturing meaningful negative projections in feature directions within steering vectors.

3.1 Steering vectors are out-of-distribution

SAEs are trained to reconstruct model activations, which have systematic differences from steering vectors. One way this out-of-distribution issue materialises, is that steering vectors have significantly

- Avoid capitals in headings, subheadings, subsubheadings (paragrahps).
- Avoid in figure titles

4 General Tips

4 General Tips

1. Use a reference manager
   - I tested Zotero and Mendeley for a month each. Found Zotero much better, though the interface is older. Zotero is more flexible and better supported.
   - Browser add-in makes your life easy.

2. Learn to use LaTeX
   - Makes everything look more professional.

3. Use LaTeX through Overleaf
   - Much easier than running LaTeX locally as has better error parsing
   - Free premium accounts at the department

4. Make multiplot figures/infographics in PowerPoint or Canva
   - You can get 1 month free for Canva pro

4 Admin Tips

1. Start writing early. It does help you form ideas
2. Give yourself a window of time at the end
3. *Extensions*
4. *Proofreading*

5

Model Essays

5 Model Essays

1. 10 model essays from students over 2020-2023

2. Marks range from 65-80

3. All very different

4. Up on Canvas and GitHub now

6 Questions