A novel NLP task and generative AI application to understand collective leadership from board text data

Daniela Valdes [1] **Rob Procter** [*1] **Gabriele Pergola** [*1] **Dimitrios Spyridonidis** [*2]

Abstract

The concept of Collective Leadership (CL, broadly speaking leadership within groups) is difficult to define and detect empirically. A promising avenue for detecting CL focuses on discursive approaches based on group interaction and 'turning points' in the discussion, where participants concur on the need for action. In this methodological paper, we present: a novel NLP-task for the detection of CL and a novel generative AI experimental architecture to support this task. To our knowledge, this research is the first to combine NLP and leadership theories with strong organisational process models. These models are linked using a formal notation applied to board text data, and cemented in a generative AI pipeline with the latest methodologies for in-context learning, reducing the need of costly manual annotation. Forthcoming research will provide an annotated dataset.

1. Introduction

Croft et al. (2022) define CL as *"The interaction of strategic ambiguity and inward- and outward-facing reification practices to maintain divergent perspectives alongside agreed collective aims, alignment, coordination of activities, and commitment to collective success."*. The literature surrounding Collective Leadership includes ample theorising but limited research on how it manifests empirically, let alone in the context of executive boards (Edwards & Bolden, 2023; Croft et al., 2022; Ospina et al., 2020; Fairhurst et al., 2020).

A promising avenue for detecting CL and connected concepts in the above definition (such as strategic ambiguity [1], reification [2] and collective work [3]) includes discursive approaches to leadership, interaction and 'turning points' in a discussion, where participants concur on the need for action (Fairhurst, 2007; Sklaveniti, 2020; Lortie et al., 2022). These degrees of reification and the distinction between divergence (or lack of coordination), collective work and collective leadership are illustrated in Figure 1.

These discursive approaches have yet to make use of Natural Language Processing (NLP) techniques to detect CL. After reviewing the NLP literature on group decision-making we identified only three articles Mayfield & Black (2019b;a; 2020) and one dataset, the Wikipedia's Article for Deletion forums (Xiao & Sitaula, 2018; Xiao & Nickerson), and no definition of an NLP task specific for detecting CL. Overall, these findings reflect that NLP (or large-scale text analytics) is hardly applied in organisational research or leadership studies (Hannigan et al., 2019).

Against this background, in this study we seek to respond to this research question: *"In the absence of a defined NLP task for the detection of CL, what is the most appropriate, generative AI architecture for identifying CL using solely board meeting textual data (board reports, minutes)?"*

This research is framed against a wider question around collective leadership (CL) in innovation adoption with a socio-technical lens within healthcare organisations (Williams & Cresswell; Krasuska et al., 2021; Cresswell et al., 2019; Hoda, 2022; Valdes et al., 2022), which serves to motivate the source data (healthcare board reports and minutes).

2. Preliminaries

2.1. Understanding how hospital executive boards work

Executive boards in public healthcare organisations are interesting places to study collective leadership. This is because

[*]Equal contribution [1]Department of Computer Science, Warwick University, UK [2]Warwick Business School, Warwick University, UK. Correspondence to: Daniela Valdes <daniela.valdes@warwick.ac.uk>.

Accepted at the 5th Latinx in AI Workshop at ICML 2024.

[1]Defined as the "deliberate use of ambiguity to accommodate competing strategic aims" Croft et al. (2022)

[2]This as opposed to divergence, can be interpreted as 'solidification of commitment'.

[3]This is defined to a situation where there is alignment and coordination of activities but no agreed direction or commitment.

Figure 1. Turning points in board discourse ordered in degrees of reification. The chart outlines how the increasing levels of reification provides a distinction between collective work and collective leadership. Low levels of reification signal divergence. After an action has been agreed and minuted, follow through signals collective leadership

of their composition, their public nature, and the language and constructs used.

Previous research in this space (Manzoor et al., 2022) has identified that these boards include nine to 11 members, roughly split between executives ('Chief' employee roles) and members of the public (non-executive directors), with the meeting led by a non-executive chairperson. Sub-board Directors author, present reports and answer questions arising from non-executives. The presence of three different stakeholder groups (executives, non-executives, directors), makes these meetings unique to establish collective leadership across layers of the organisational structure.

Given their public sector nature, members of the public can attend and observe these meetings, and reports and papers are matters of public record. While meeting observations are traditionally used for leadership research (Croft et al., 2022; Sklaveniti, 2020; Lortie et al., 2022; Manzoor et al., 2022), there is limited analysis of the content of the reports and discussions provided (Watkins et al., 2008), making the reports and minutes of these meetings an untapped source of data for computational scientists.

Although beyond the scope of this paper, it is worth noting the linguistic features and availability of this data are complex for data scientists. A forthcoming dataset will be one of our contributions of our wider study ((Valdes et al., 2024)).

2.2. A worked example of collective leadership in board data

What does this mean in practice? The degrees of reification and the distinction between collective work and collective leadership were illustrated in Figure 1. As noted in that figure, we will focus on the ability to identify allocated actions which are followed through over time. This is illustrated in Figure 2. In the first section of the figure, we see the section

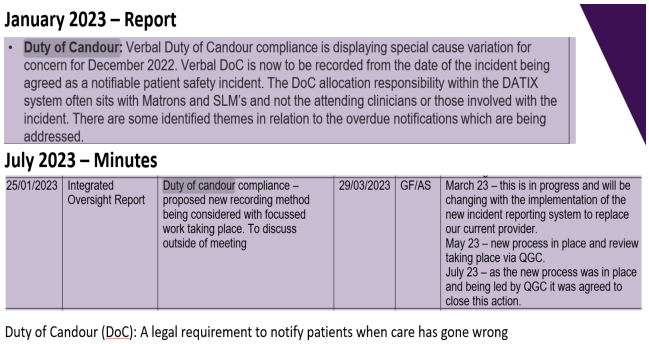

Figure 2. Exampleof collective leadership in board text data. These two extracts provide an example of the linguistic features we want the model to focus on: identifying an action allocated to an individual or committee, and how this specific action has been followed through over time.

of the report in January 2023 discussing the element related to duty of candor compliance. This generated an action for members GF/AS, as can be seen in the extract of the July action log. This action was effectively followed through until July, when it was agreed to be closed as considered complete.

Considering that expenditure in healthcare services is over £70bn in England, there is a strong public interest in providing automated means to verify leadership cohesion in these organisations. We envisage these principles could be extended to other large organisations, outside of healthcare. In summary, our NLP task is a novel, useful, challenging task that has impact in the real world.

In the next section, we discuss more in detail how we define the CL NLP task.

3. A novel NLP task to identifying collective leadership from text data

Introducing our notation, l represents an element of the set of board decision-making labels; s,f represent paragraphs of the minutes and reports respectively; i represents a particular hospital; t, τ represent different meetings (or moments in time), and the symbol $\sim$ denotes similarity.

In this section we formally introduce the discussion labels l, in particular the 'Accept Action' label to detect collective work and collective leadership. We link this with the organisational research contexts mentioned in section 2.1 above.

To inform our NLP approach, we translate the concept of CL into an executive board space of NHS hospitals by focusing on particular 'turning points' in the discussion, where participants formally agree on the need for change through a minuted action. Following (Croft et al., 2022) and as illustrated in 1 above, we posit that to detect collective leadership from executive board text requires the fulfilment of two conditions:

1. Collective work ($C_w(t,\tau)$) or joint understanding across time can be detected by comparing the semantic similarity between sections of minutes and reports for a particular meeting and over time across meetings. We consider that most types of board actions (in the textual form of a *decision-making label* l) can signal collective work, provided that we see some commonality/similarity between reports and discussions over time.

Equation 1 shows that there is collective work in the form of joint understanding over time, if we can identify sufficiently similar text within board minutes and reports, for any type of board discussion labels (as long as they are different from 'Accept Action'). The concept is formulated below.

$$\forall t,\tau \in 1,...,T, C_w \iff (l_i^t \neq `AcceptAction') \wedge \\ \wedge (s_i^t \sim f_i^t \sim s_i^\tau \sim s_i^\tau) \wedge (t \neq \tau) \text{ where } C_w \tag{1}$$

2. Reification over strategic ambiguity (R(t) happens when there is a clear, agreed action for a nominated individual, signalling there is enough 'solidification of commitment' at the executive level to merit a change in direction. We will thus have a section of the minutes labelled with 'Accept action'.

$$R(t) \iff l_i^t = `AcceptAction' \tag{2}$$

Collective leadership ($CL(t,\tau_1,\tau_2)$) takes place when we see an 'Accept Action' label as part of a discussion in the minutes, provided that features of that discussion will have some follow-up over time (in future), and there has been some discussion about it (contemporaneously or in the past).

Applying the label 'Accept Action' at time t for a section of the minutes s_i^t in isolation does not reflect CL. It does so only if we see (i) *sustained commitment over time* through other minutes or reports with a similar topic in future, ($s_i^{\tau_2}$,$f_i^{\tau_2}$) and (ii) *evidence of previous collective work* $C_w(\tau_1)$, which means it has also been raised previously in minutes or reports).

Equation 3 shows there must be at least two points in time (in past -τ_1 and in future -τ_2) where the language model finds semantic similarity compared to the text (s_i^t) in time t which has been classified with an 'Accept Action' label.

This is formalised in equation 3 below.

$$CL(t,\tau_1,\tau_2) \\ \iff \exists \text{ t}, \tau_1,\tau_2, \in 1,\ldots,\tau_1,\ldots,t \ \ldots,\tau_2,\ldots,T \ \wedge \\ \wedge R(t) \wedge C_w(\tau_1,t) \wedge C_w(\tau_2,t) \tag{3}$$

Once we have identified the relevant section of the minutes dealing with a particular action, we verify the condition of collective work over time. We do this by identifying that a similar text which can be found in other minutes and reports at other points in time (in future).

4. A novel generative AI architecture for identifying CL in board text data

We propose use large language models and generative AI to inform two separate, well-known NLP tasks: one, a text classification task which allows the large language model to identify when an action has been recorded as such within the minutes (the 'Accept action' label). Then, a semantic similarity task, which allows the model to identify when a similar topic has been raised in previous reports, and has been discussed in future meetings (signalling collective leadership).

Figure 3 outlines the proposed architecture of the language model. Within the architecture, we have considered various potential text representations and prompting approaches as part of our experimental design. Our experimental design considers 12 (3x4) architectures as outlined below. In our notation σ is a parameter that denotes a quantitative threshold for semantic similarity (such as Dice Coefficient or Jaccard Index (Peinelt, 2021)). As part of our experimental setting, we will test various levels of σ.

- Input: This includes a (forthcoming) dataset containing our corpus of board-level documents (reports and minutes, split in paragraphs f and s respectively) for each hospital h for the period 2017-2023.

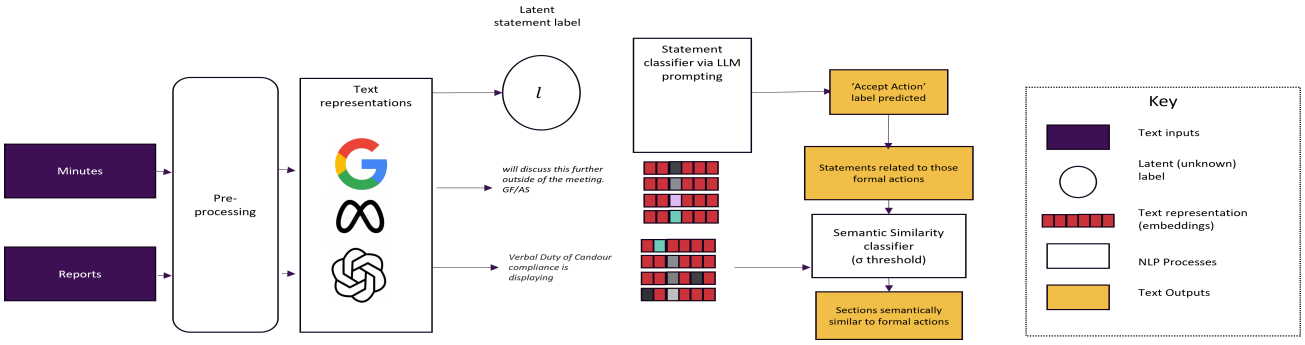

Figure 3. Proposed NLP architecture. The model receives text from the minutes and reports to create embeddings. In the first instance, a text-classification tool processes the minutes to identify formal actions ('Accept Action' label). Sections of the minutes classified with that label go through a semantic similarity classifier to identify other similar texts in future board reports/minutes.

- **Training:** We will train the classification model using a manually labelled subset of the dataset drawn from minutes from a selection of hospitals from our wider sample, splitting the dataset in 80\10\10 proportion. We will aim to have at least 10 examples of each label as per Brown et al. (Brown et al., 2020).
- **Text representations and AI/Language Models:** We will consider three different text representations using GPT-4 (by Open AI)(OpenAI, 2023), LlaMa 2 (by Meta)(noa) and BERT (by Google) (Devlin et al., 2019), which will be used for text classification to identify the latent 'accept action' label.
- **Prompts:** We will consider four different prompting methods: zero/few-shot (Brownlee, 2018), chain of thought prompting (Wei et al., 2023), chain of density (Adams et al., 2023).
- **Semantic Similarity**. As this is a standard NLP task, we propose to use a single architecture, tBERT (Peinelt, 2021).
- **Evaluation**. When creating the overall Collective Leadership NLP task we face an evaluation challenge as there is no established 'ground truth', something to benchmark the model against. In this case, the NLP literature suggests a combination of quantitative (balanced accuracy and micro- macro-F1), qualitative and human-based evaluation techniques drawing from computational grounded theory (CGT) (Mayfield & Black, 2019a; Nelson, 2020), so selected CL passages will be subject to human deep-reading.

5. Discussion

In this short paper we have established a working definition of collective leadership to motivate a novel NLP task. The strengths and limitations of our approach are outlined below.

5.1. Strengths

a. Theoretically robust approach, drawing on strong process models (Croft et al., 2022; Denis et al., 2011; Sklaveniti, 2020; Lortie et al., 2022) in organizational and collective leadership research, motivated through rigorous mathematical notation and rooted socio-linguistic leadership literature (Fairhurst, 2007).

b. Building upon the use of the latest generative AI for text classification, employing natural language prompting.

c. Empirically grounded, utilising publicly available empirical text data from hospital boards.

5.2. Limitations

We explore potential limitations arising from various biases (methodological, data, researcher bias) as well as accuracy of pre-trained language models.

a. **Methodological biases and errors.** These might emerge through the pre-processing (encoding) of textual data. We seek to minimise these biases by testing various encoding approaches and evaluating their performance as outlined in the evaluation section. We are aware quantitative approaches are not bias-free (particularly given the use of natural language to 'prompt' the AI towards a particular text classification) (Tschisgale et al., 2023). We mitigate these biases by approaching the analysis iteratively in a cyclical manner, alternating between human and computational tasks.

b. **Data biases.** Minutes, committee documents and routine reports are classified as 'reportative' (Heller, 2023) sources containing factual, historical information with limitations arising from 'authorship, bias and power' (Heller, 2023). We mitigate this by asking research

questions with a focus on organisational practices, including contextual analysis and a representative sample to support triangulation. To reduce data biases we propose to request additional documentation from relevant hospitals to further inform leadership actions. LLMs, as repositories of language data, include social biases around gender, race, religion and social constructs (Liang et al., 2021).

c. **Researcher bias**. Any research design reflects the researcher's perspective, which is shaped by their own beliefs and the scientific community they belong to (Kaur & Kumar, 2021). The use of Computational Grounded Theory (CGT) as part of the wider researh study includes a subjective analysis and coding of results, which might reflect researcher bias and might be difficult to reproduce. We also mitigate researcher bias through reflexivity, intended as a "mutual shaping between researcher and research" (Attia & Edge, 2017) to support the researcher's developmental journey as a Computer Scientist.

d. **Accuracy of pre-trained language models.** Our approach intends to build upon pre-trained LLMs which are domain-agnostic. While pre-trained models using domain-specific, pre-annotated data might be able to achieve higher levels of accuracy and performance, there is a large cost annotating this data (Tschisgale et al., 2023). Our training is limited to the labelling of a small section of out-of-sample board reports to achieve a few examples of the different types of 'discussion labels' to classify sections of the minutes ((Valdes et al., 2024)). We have noted the limited data availability of board text data for other organisations and industries, and future research could identify whether our proposed label/action taxonomy applies to other types of boards in the public or private sector. Further avenues can also consider argumentational analysis of the decision-making labels as to justify a particular course of action (NLP tasks of argument mining). As part of our model optimisation, we have proposed an experimental approach and a set of metrics to find the best combination of word embeddings, language models and algorithms for the detection of CL.

References

Llama 2: Open Foundation and Fine-Tuned Chat Models | Research - AI at Meta. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., and Elhadad, N. From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting, September 2023.

Attia, M. and Edge, J. Be(com)ing a reflexive researcher: A developmental approach to research methodology. *Open Review of Educational Research*, 4(1):33–45, January 2017. ISSN null. doi: 10.1080/23265507.2017.1300068.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners, July 2020.

Brownlee, J. A Gentle Introduction to k-fold Cross-Validation, May 2018.

Cresswell, K., Sheikh, A., Krasuska, M., Heeney, C., Franklin, B. D., Lane, W., Mozaffar, H., Mason, K., Eason, S., Hinder, S., Potts, H. W. W., and Williams, R. Reconceptualising the digital maturity of health systems. *The Lancet Digital Health*, 1(5):e200–e201, September 2019. ISSN 2589-7500. doi: 10.1016/S2589-7500(19)30083-4.

Croft, C., McGivern, G., Currie, G., Lockett, A., and Spyridonidis, D. Unified Divergence and the Development of Collective Leadership. *Journal of Management Studies*, 59(2):460–488, 2022. ISSN 1467-6486. doi: 10.1111/joms.12744.

Denis, J.-L., Dompierre, G., Langley, A., and Rouleau, L. Escalating Indecision: Between Reification and Strategic Ambiguity. *Organization Science*, 22(1):225–244, February 2011. ISSN 1047-7039. doi: 10.1287/orsc.1090.0501.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

Edwards, G. and Bolden, R. Why is collective leadership so elusive? *Leadership*, 19(2):167–182, April 2023. ISSN 1742-7150. doi: 10.1177/17427150221128357.

Fairhurst, G. T. *Discursive Leadership: In Conversation with Leadership Psychology*. SAGE Publications, Incorporated, Thousand Oaks, UNITED STATES, 2007. ISBN 978-1-4522-6672-5.

Fairhurst, G. T., Jackson, B., Foldy, E. G., and Ospina, S. M. Studying collective leadership: The road ahead. *Human Relations*, 73(4):598–614, April 2020. ISSN 0018-7267. doi: 10.1177/0018726719898736.

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., and Jennings, P. D. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management*

Annals, 13(2):586–632, July 2019. ISSN 1941-6520. doi: 10.5465/annals.2017.0099.

Heller, M. Rethinking Historical Methods in Organization Studies: Organizational Source Criticism. *Organization Studies*, 44(6):987–1002, June 2023. ISSN 0170-8406. doi: 10.1177/01708406231156978.

Hoda, R. Socio-Technical Grounded Theory for Software Engineering. *IEEE Transactions on Software Engineering*, 48(10):3808–3832, October 2022. ISSN 0098-5589, 1939-3520, 2326-3881. doi: 10.1109/TSE.2021.3106280.

Kaur, M. and Kumar, R. Mixed Methods in Global Health Research. In *Handbook of Global Health*, pp. 239–260. Springer, Cham, 2021. doi: 10.1007/978-3-030-45009-0_11.

Krasuska, M., Williams, R., Sheikh, A., Franklin, B., Hinder, S., TheNguyen, H., Lane, W., Mozaffar, H., Mason, K., Eason, S., Potts, H., and Cresswell, K. Driving digital health transformation in hospitals: A formative qualitative evaluation of the English Global Digital Exemplar programme. *BMJ Health & Care Informatics*, 28(1):e100429, December 2021. ISSN 2632-1009. doi: 10.1136/bmjhci-2021-100429.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6565–6576. PMLR, July 2021.

Lortie, J., Cabantous, L., and Sardais, C. How Leadership Moments are Enacted within a Strict Hierarchy: The case of kitchen brigades in haute cuisine restaurants. *Organization Studies*, pp. 01708406221134225, October 2022. ISSN 0170-8406. doi: 10.1177/01708406221134225.

Manzoor, H., Nocker, M., and Boncori, I. The performativity and politics of emotions in NHS boards. *Culture and Organization*, 28(6):509–527, November 2022. ISSN 1475-9551. doi: 10.1080/14759551.2022.2105337.

Mayfield, E. and Black, A. Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pp. 65–77, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-2108.

Mayfield, E. and Black, A. W. Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, November 2019b. ISSN 2573-0142. doi: 10.1145/3359308.

Mayfield, E. and Black, A. W. Should You Fine-Tune BERT for Automated Essay Scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 151–162, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.15.

Nelson, L. K. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 49(1):3–42, February 2020. ISSN 0049-1241. doi: 10.1177/0049124117729703.

OpenAI. GPT-4 Technical Report, March 2023.

Ospina, S. M., Foldy, E. G., Fairhurst, G. T., and Jackson, B. Collective dimensions of leadership: Connecting theory and method. *Human Relations*, 2020. doi: 10.1177/0018726719899714.

Peinelt, N. *Detecting Semantic Similarity : Biases, Evaluation and Models*. PhD thesis, University of Warwick, January 2021.

Sklaveniti, C. Moments that connect: Turning points and the becoming of leadership. *Human Relations*, 73(4):544–571, April 2020. ISSN 0018-7267. doi: 10.1177/0018726719895812.

Tschisgale, P., Wulff, P., and Kubsch, M. Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, 19(2):020123, September 2023. doi: 10.1103/PhysRevPhysEducRes.19.020123.

Valdes, D., Hijazi, G., Shanker, A., Mensah, D., Bockaire, T., Lazar, I., Ibrahim, A., Zolfagharinia, H., Procter, R., Spencer, R., and Dale, J. Evidence on the sustainability of telemedicine in outpatient and primary care during the first two years of the COVID-19 pandemic: A global scoping review (Preprint). Preprint, Journal of Medical Internet Research, December 2022.

Valdes, D., Procter, R., Pergola, G., and Spyridonidis, D. A summarised datasheet for the Noren dataset (forthcoming). 2024.

Watkins, M., Jones, R., Lindsey, L., and Sheaff, R. The clinical content of NHS trust board meetings: An initial exploration. *Journal of Nursing Management*, 16(6):707–715, 2008. ISSN 1365-2834. doi: 10.1111/j.1365-2834.2008.00928.x.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023.

Williams, R. and Cresswell, K. 30 Month Report - Global Digital Exemplar Evaluation Pro- gramme,. Technical report, The University of Edinburgh.

Xiao, L. and Nickerson, J. V. Imperatives in Past Online Discussions: Another Helpful Source for Community Newcomers?

Xiao, L. and Sitaula, N. Sentiments in Wikipedia Articles for Deletion Discussions. In Chowdhury, G., McLeod, J., Gillet, V., and Willett, P. (eds.), *Transforming Digital Worlds*, Lecture Notes in Computer Science, pp. 81–86, Cham, 2018. Springer International Publishing. ISBN 978-3-319-78105-1. doi: 10.1007/978-3-319-78105-1_10.