III Międzynarodowa Konferencja Naukowa
Instytutu Zarządzania UEP
ZRÓWNOWAŻONE ZARZĄDZANIE ORGANIZACJĄ
3rd International Scientific Conference
of the PUEB Institute of Management
SUSTAINABLE ORGANIZATION MANAGEMENT
Zrównoważone Zarządzanie Organizacją
Instytut Zarządzania
UNIWERSYTET EKONOMICZNY W POZNANIU
SAMSUNG
21–22 października / October 2026

Replicating human rhetorical coding of ESG reports with language models

Mariusz Hofman
Maria Curie-Skłodowska University in Lublin

Introduction and research aim

As Higgins and Walker (2012) show, ESG reports persuade through logos (facts and figures), ethos (the sender's credibility) and pathos (the reader's emotions; Plutchik, 2001). Coding them manually at the sentence level is labour-intensive and limits the scale of research, while language models can outperform crowd workers in text annotation (Gilardi et al., 2023).

Aim: to test whether language models reproduce the coders' decisions when coding ethos, pathos and logos.

Research questions:
RQ1. How closely does the model reproduce the agreed coding?
RQ2. Does an open local model match a closed model?
RQ3. Does the JEV protocol with calibrated confidence preserve coding quality?

Research methods

Current status: two ESG reports (BP and Mazda), about 2,250 sentences in total, with the agreed coding of two coders.

Configurations (same coding rules):
- closed: gpt-5.6-luna with guardrails (corrective patterns),
- local: open qwen3.8:27b on a university server, without thinking mode, with guardrails,
- JEV: the JEV-27B protocol (AutoTrust AI, 2026) on the same qwen3.8:27b, calibrated to the coders' decisions.

Measure: min-F1, i.e. the F1 of the weaker of the pathos and ethos classes; differences below 3 points (noise: 1.6 points) are inconclusive.

Study plan: calibration on four reports coded by the coders (BP, Alior Bank, Mazda, Amazon), followed by machine coding of 100 ESG reports.

Results and discussion

- Coding quality depends mainly on the coding rules.
- Once the rules were derived from the coders' practice in both reports, the mean score of the local configuration rose from 52.8 to 64.2 min-F1 (RQ1). It outperformed the closed configuration (57.8) at no cost and without sending data outside the university (RQ2).
- JEV reached 59.4. It ranks sentences well by confidence, but a decision threshold set on one report transfers poorly to another (RQ3).
- Many of the remaining errors, especially for ethos, involve sentence types that the coders themselves coded differently in the two reports, suggesting that further gains depend more on harmonising the coding rules than on a stronger model.

This research was carried out within the project “Does colouring in ESG reports pay off? Investigating the use and effects of rhetoric in mandatory ESG reporting”, funded by the National Science Centre (Poland), grant no. 2025/57/B/HS4/00176.

Conclusions

The local configuration with rules derived from the coders' practice achieves a mean min-F1 of 64.2 at no cost, which makes it possible to move on to analysing rhetoric in 100 ESG reports.

Limitations: the rules were developed on the same two reports, so the results need to be confirmed on the Alior Bank and Amazon reports.

References

Higgins, C., Walker, R., 2012, Ethos, logos, pathos: Strategies of persuasion in social/environmental reports, Accounting Forum, 36(3), 194–208.
Gilardi, F., Alizadeh, M., Kubli, M., 2023, ChatGPT outperforms crowd workers for text-annotation tasks, PNAS, 120(30), e2305016120.
Plutchik, R., 2001, The nature of emotions, American Scientist, 89(4), 344–350.
AutoTrust AI, 2026, JEV-27B [language model], Hugging Face, https://huggingface.co/autotrust/JEV-27B.