# Training-Free Control of Moral Framing Preservation in Automated News Summarization

Project Assignment

## Project Description

News articles are more than collections of facts: journalists *frame* events, emphasizing some aspects and downplaying others to shape how the audience perceives an issue [1]. One important aspect of framing is the choice to write in, or quote verbatim, morally charged language rather than neutral terms. When an LLM summarizes an article, this framing can be silently lost. For example, "the NGOs criticized the plan" may be a factually accurate summary of an article's section, but it may erase the journalist's intention to report the NGOs' strong condemnation.

Liscio et al. [4] present the first study of moral frame preservation in AI-generated news summaries. They use the EMONA dataset [3], which contains word-level annotations of moral-ladenness through the lens of the Moral Foundations Theory (MFT) [2]. They show that prompting strategies that ask the model to preserve moral-laden words improve preservation, but that misguided prompting can lead summarizers to *add* moral framing to neutral parts of the article.

However, their methods operate on the prompt. In this assignment, you will instead intervene on the model's internal representations, in the spirit of training-free alignment [5]: steering behavior at inference time without updating weights. You will use contrastive activation addition (CAA) [6], where a direction in the residual stream is estimated from the average difference between activations on contrasting examples, and is then added to or subtracted from the hidden states during generation.

Here the target is moral-ladenness in summarization. You will build a direction from EMONA sentences that contain annotated moral-laden words and sentences that do not, and use it to push a summarizer toward preserving moral language (positive steering) or neutralizing it (negative steering). This is a research-style project with an uncertain outcome: a direction learned from short sentences may capture correlated properties (sentiment, topic) rather than moral-ladenness, and may not transfer to article-level summarization. Finding out is part of the task, and carefully analyzed negative results are fully acceptable.

**RQ1: Steerability**
Does adding (subtracting) the moral-ladenness direction increase (decrease) the preservation of annotated moral-laden words in summaries of unseen articles, compared to an unsteered baseline? How does this depend on the layer and the steering strength?

**RQ2: Foundation specificity**
Which MFT elements are most preserved and which most lost under positive and negative steering? Is the effect uniform or concentrated in some elements, and is it symmetric between directions?

**RQ3: Cost**
What does steering cost in summary quality, and is there a setting with a meaningful change in moral preservation at negligible cost?

## Experimental Setup

1. **Model.** Choose an instruction-tuned open model (in the 1–8B parameter range).
2. **Data and baseline.** Obtain EMONA [3] and split it by article into fit, validation and test sets (possibly using the stratified 60-article test set of [4]). Generate unsteered summaries with the *Plain* prompt of [4] (200-word limit) as your baseline.
3. **Contrastive data.** From the fit articles, build two datasets of sentences with and without annotated moral-laden words, respectively. Balance for obvious confounds such as length and article. Inspect a handful of sentences by hand to note what else differs.
4. **Steering direction.** For each sentence in the two datasets created in the previous point, perform an inference pass and record the model's activations at each layer. Compute the average activations obtained with the sentences annotated with moral-laden words and the average activations obtained with the sentences without moral-laden words. Calculate the difference between those two averages to obtain a steering vector.
5. **Steering sweep.** Add the direction (positive) or subtract it (negative) while generating summaries, over a range of steering strengths ($\alpha$) and layers. Sweep on the validation set to choose the best-performing layers and steering strengths.
6. **Evaluation.** Measure overall moral preservation with the Moral Count and Moral Divergence metrics of [4], and compute the preservation rate of each MFT element for baseline, positive and negative steering. Measure summary quality with at least one reference-free metric taken from [4]. Control for length, since longer summaries tend to preserve more moral words, and use paired statistical tests against the baseline.
7. **Trade-offs and qualitative analysis.** Plot moral preservation against each quality metric across steering strengths, and identify where quality breaks down. Read summaries from both directions, categorize them (e.g., with the expert labels of [4]: framing loss, addition, modification, quote omission), and check explicitly whether positive steering adds moral language not present in the article.

## References


[1] R. M. Entman. Framing: Toward clarification of a fractured paradigm. In *Journal of Communication*, 1993.

[2] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, 2013.

[3] Y. Lei, M. M. M. Miah, A. Qamar, S. R. Reddy, J. Tong, H. Xu, and R. Huang. EMONA: Event-level moral opinions in news articles. In *Proceedings of NAACL 2024*.

[4] E. Liscio, M. Lorandi, and P. K. Murukannaiah. News is more than a collection of facts: Moral frame preserving news summarization. In *COLM 2025*.

[5] B. Pan, Y. Li, W. Zhang, W. Lu, M. Xu, S. Zhou, Y. Zhu, M. Zhong, and T. Qian. A survey on training-free alignment of large language models. In *Findings of EMNLP 2025*.

[6] J. Wehner, S. Abdelnabi, D. Tan, D. Krueger, and M. Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. In *Transactions on Machine Learning Research*, 2025.