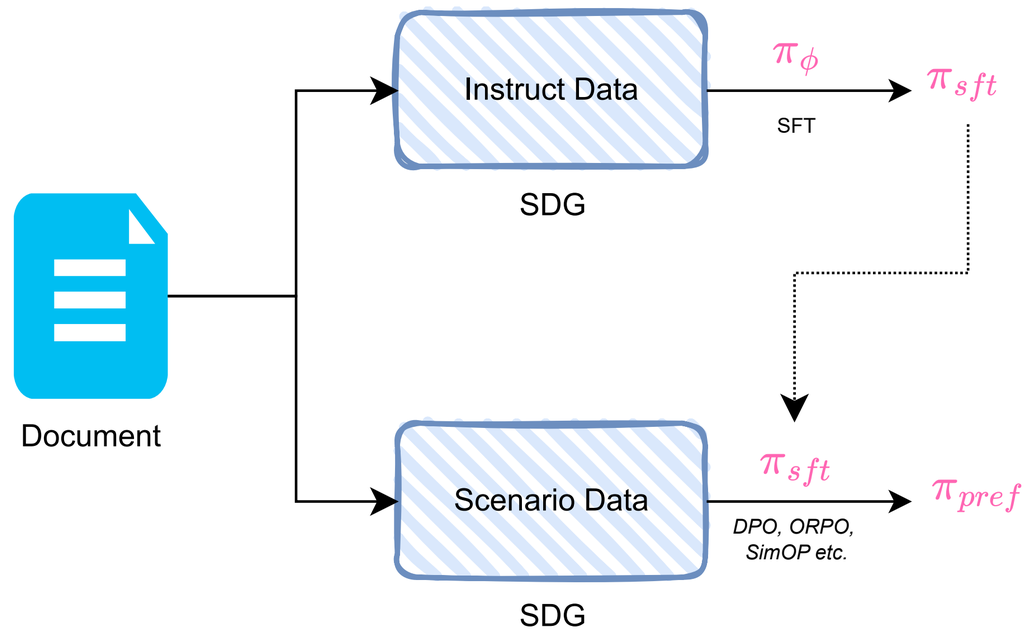
Value Alignment From Unstructured Text

IBM Research

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri,
Manish Nagireddy, Pierre Dognin, Kush R. Varshney
IBM Research, New York

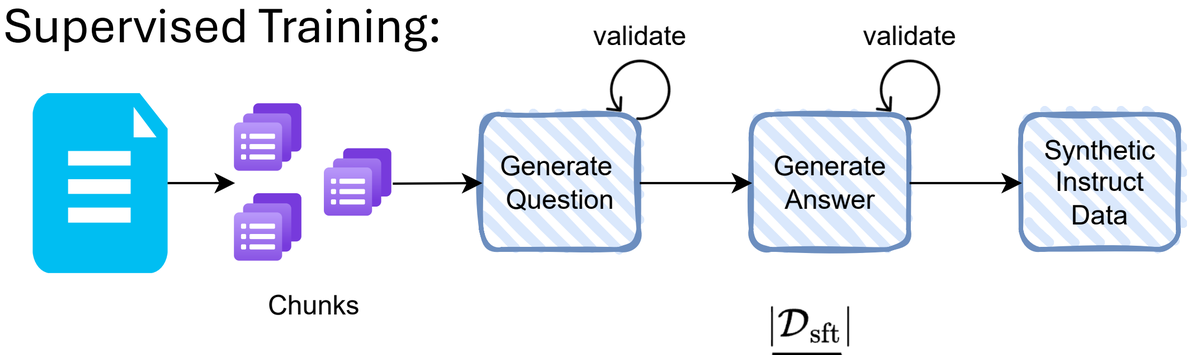
Overview

- Alignment methods depend on high-quality handcrafted instruction and preference data.
- The goal is to build systems that adhere to the contextualized values embedded within ***unstructured text***
- ***Contextualized values*** can stem from individuals, communities, companies, and other sources.
- An end-to-end methodology to effectively align LLMs with values that are ***implicitly*** and/or ***explicitly*** engraved in the unstructured text.



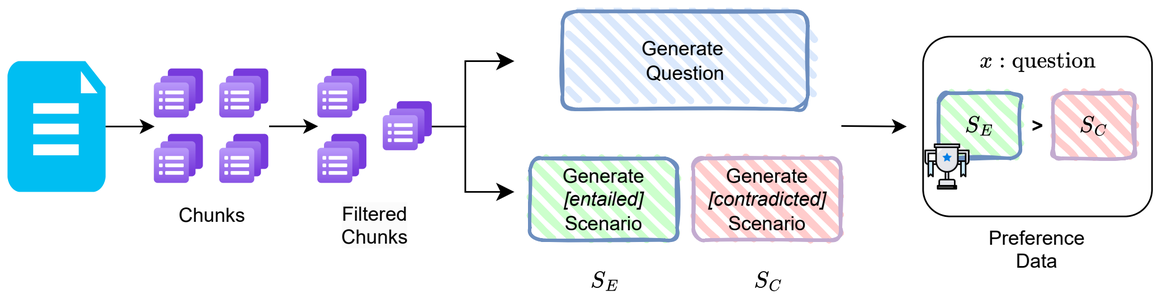
Synthetic Data Generation

Supervised Training:



$$\pi_{\text{sft}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{|\mathcal{D}_{\text{sft}}|} -\log \pi_\theta(y_i|x_i)$$

Preference Optimization:



$$\pi_{\text{pref}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{|\mathcal{D}_{\text{pref}}|} -\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_{iw} \mid x_i)}{\pi_{\text{sft}}(y_{iw} \mid x_i)} - \beta \log \frac{\pi_\theta(y_{il} \mid x_i)}{\pi_{\text{sft}}(y_{il} \mid x_i)}\right)\right]$$

Use Cases

Efficiency and effectiveness study using two use cases:

- ***Business Conduct Guidelines***
  - Corporate business guideline that provides set of principles & rules for employees
  - 46 pages covering values like conflict of interest, discrimination, harassment, transparency, etc.
- ***UDHR***
  - Universal Declaration of Human Rights document by the UN
  - Sets out fundamental human rights and broad range of civil, social, cultural, and economical rights
  - UDHR is one of the sources for Constitutional AI principles.

Results

Models:

- seed : *mistral-7b-instruct-v0.2*
- teacher: mixtral-8x7B-Instruct-v0.1

- Our method outperforms related competitive methods when evaluated using automatic metric & win-rates
- ⚠ Integrating RAG on aligned model hampers performance
- ‼ UDHR aligned models improves general safety

BCG

Model	RAG	BLEU	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum	BertScore	winrate
`c-fine-tuned`	✓	26.067	0.555	0.336	0.409	0.427	0.918	0.524±0.08
our method								
+ SFT π_{sft}	✓	32.744	0.606	0.434	0.494	0.507	0.929	0.389±0.10
+ DPO π_{pref}	✓	32.693	0.606	0.434	0.494	0.507	0.929	0.390±0.10
our method								
+ SFT π_{sft}	✗	36.667	0.628	0.453	0.517	0.536	0.918	0.603±0.07
+ DPO π_{pref}	✗	38.528	0.633	0.457	0.521	0.540	0.932	0.615±0.06

UDHR

Model	RAG	BLEU	Rouge1	Rouge2	Rogue-L	Rouge-Lsum	BertScore	winrate
`c-fine-tuned`	✓	22.946	0.528	0.311	0.376	0.399	0.911	0.497±0.05
our method								
+ SFT π_{sft}	✓	31.333	0.604	0.422	0.480	0.502	0.926	0.492±0.09
+ DPO π_{pref}	✓	31.228	0.604	0.423	0.480	0.502	0.926	0.478±0.09
our method								
+ SFT π_{sft}	✗	35.554	0.629	0.449	0.508	0.536	0.929	0.649±0.06
+ DPO π_{pref}	✗	35.689	0.630	0.451	0.509	0.537	0.929	0.640±0.07

References:
[1] : Sorensan et al., "*A Roadmap to Pluralistic Alignment*", in ICML 2024
[2] : S. Achintalwar et al., "*Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations*," in IEEE Internet Computing, doi: 10.1109/MIC.2024.3453671.