THEME ARTICLE: CIVILIZING AND HUMANIZING AI

Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations

Swapnaja Achintalwar, Ioana Baldini, Djallel Bouneffouf, Joan Byamugisha, Maria Chang, Pierre Dognin, Eitan Farchi, Ndivhuwo Makondo, Aleksandra Mojsilović, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Inkit Padhi, Orna Raz, Jesus Rios, Prasanna Sattigeri, Moninder Singh, Siphiwe A. Thwala, Rosario A. Uceda-Sosa, and Kush R. Varshney, *IBM Research*

The alignment of large language models is usually done by model providers to add or control behaviors that are common or universally understood across use cases and contexts. By contrast, in this article, we present an approach and architecture that empowers application developers to tune a model to their particular values, social norms, laws, and other regulations and orchestrate between potentially conflicting requirements in context. We lay out three main components of such an Alignment Studio architecture: Framers, Instructors, and Auditors, which work in concert to control the behavior of a language model. We illustrate this approach with a running example of aligning a company's internal-facing enterprise chatbot to its business conduct guidelines.

Pretrained large language models (LLMs) are usually tuned by model providers to endow them with different abilities, such as the ability to follow instructions and conduct helpful conversations with the user. Many model providers perform further tuning, known as *alignment*, to make the LLM helpful and harmless according to their definitions of helpfulness and harmlessness. These steps of "civilizing" and "humanizing" the LLM are decisive in controlling the model's behavior, more so than the pretraining of the base model. The harms that model providers aim to prevent are common ones found in risk taxonomies, such as hate, malice, exclusion, profanity, and toxicity, which have existing benchmarks and evaluation datasets.

Nevertheless, we do not believe that such alignment to common concerns can ever be comprehensive, and we do not believe that all dimensions of alignment are always necessarily desirable. Context matters. Every industry, sector, jurisdiction, culture, and use case has its own unique and *particular* desired behaviors that are *not* captured in a *common* taxonomy. The examples are numerous. In a medical application, developers may not want an LLM to treat names of body parts as profanity. In a customer complaint processing application whose inputs are laced with offensive language, developers may want the system to continue to operate. A grocery store's chatbot may have an extra requirement to refrain from mentioning poisonous food items, and a bank's chatbot may have an extra requirement to refrain from mentioning competitor's brands or products. Laws may require certain LLM behaviors, like the one in China that requires all generative content to reflect core socialist values. An organization may have a style guide for the LLM's tone and personality to which it must adhere. Companies or professions may have guidelines that specify the business conduct to be respected. All of these examples are valid desired behaviors depending on the context, but they would not show up in the alignment done by model providers for common concerns.[a]

There are many sources of regulations, not only laws but also social norms, market demands, and technological constraints.[1] The associated requirements can be quite unique and contextual based on the use case. As such, they will not have existing benchmarks with which to evaluate the LLM. Additionally, different regulations may be competing or conflicting.

1089-7801 © 2024 IEEE
Digital Object Identifier 10.1109/MIC.2024.3453671
Date of publication 5 September 2024; date of current version 25 November 2024.

[a] We, as authors, may not agree with aligning an LLM to all of the listed examples. But that is the point: our personal values are not universal and we should not impose them on end communities. We have given the examples in the spirit of providing a broad aperture on the possibilities.

In contrast to high-level general statements like Bai et al.'s[2] "Do NOT hoose responses that exhibit toxicity, racism, sexism, or any other form of physical or social harm," particular regulations may be quite detailed. At IBM, where we work, we have detailed business conduct guidelines (BCGs) and will use these as a running example of a particular set of regulations with which to align LLM behavior.

Adherence to various particular contextual regulations has many business benefits, from better serving customers to avoiding prosecution. Perhaps the biggest benefit is making an LLM authentic to the values of the model deployer and the community of end users. It is a form of personalization or customization,[3] a kind of steerable pluralism,[4] and a method of decoloniality that dismantles the power of model providers and empowers communities to have a say in what is "civil" and "human."[5] Importantly, application developers can only further align LLMs beyond the common alignment done by model providers if the models are *open*. Furthermore, alignment techniques must not be too costly or burdensome so that they extend beyond the means of application developers.

The customization of the LLM's behavior to nonuniversal values and requirements calls for tooling that we name *Alignment Studio*. The starting point is a set of regulations given in a natural language policy document, which could be a law, an industry standard, or a corporate guideline that has already been deliberated upon and adopted.[b] The tooling permits a principled, transparent, auditable, and deliberate approach to alignment.

As illustrated in Figure 1, the "Framers" component of Alignment Studio produces the necessary data for instilling desired behaviors and evaluating whether we were successful in doing so. The "Instructors" component uses these data to fine-tune the model, and additionally allows for the orchestration of competing values or regulations. The "Auditors" component is responsible for the evaluation of the alignment. Framers, Instructors, and Auditors form a continuous cycle of development. A test-driven development approach could also, for instance, begin with Auditors. A representative software architecture for Alignment Studio, starting with policy documents, is illustrated in Figure 2.

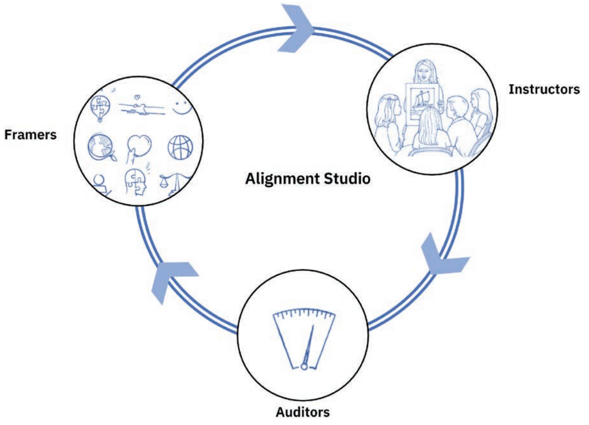


FIGURE 1. A stylized depiction of Alignment Studio with its three components: Framers, Instructors, and Auditors.

Broadly, the purpose of aligning an LLM to a particular set of regulations is to control or govern its behavior, which is the main system-level functionality provided by Alignment Studio. Our work is distinct from the closest similar methodologies like Constitutional AI[2] and Self-Instruct[6] because of its emphasis on automation in Framers that begins with unstructured text, thereby facilitating the specification of alignment that enterprises need in particular. Moreover, the approach uniquely includes ontological reasoning, which addresses the fact that because real-world regulations are often quite specific in their intent (e.g., an employee may not work for a competitor), they often require knowledge beyond what is given in the document (e.g., who constitutes a competitor in the aforementioned case).

Some of our work reuses known components from the literature on alignment, especially the basic technical algorithms for fine-tuning and preference optimization, but our contribution is mainly at the system level. Nevertheless, we contribute to pluralistic and contextual alignment technologies[7] as another important dimension of artificial intelligence (AI) governance provided by Alignment Studio is transparently and controllably choosing among possibly conflicting behaviors. Lazar[8] summarizes the need for such governance that is missing from the literature: "Steering LLM behaviour is actually a matter of governing their end users, developing algorithmic protections to prevent misuse. If this algorithmic governance depends on inscrutable trade-offs made by an LLM, over which we have no explicit or direct control, then that governing power is prima facie illegitimate and unjustified." This article is not a survey of alignment techniques.

RUNNING EXAMPLE

In this article, we use the following running example to illustrate the Alignment Studio: An LLM is infused into an application for IBM employees to ask general questions, receive advice, and receive suggestions. The LLM is aligned to IBM's corporate policies

[b] In future work, we will expand to other forms of values specifications.

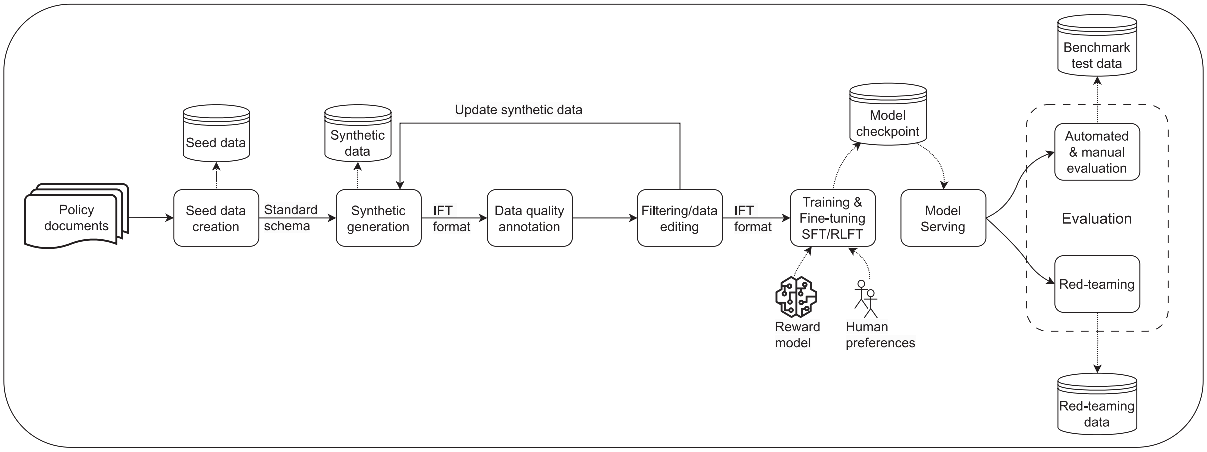


FIGURE 2. A realization of the Alignment Studio software architecture, starting with policy documents. End-to-end software testing and documentation is recommended, but implementations need not include all components. SFT: supervised fine-tuning; RLFT: reinforcement learning fine-tuning; IFT: instruction fine-tuning.

documented in the IBM BCGs,[c] which is a 46-page document with roughly 11,500 words divided into eight sections with 78 paragraphs. The content is expressed in different forms, such as *topic-paragraph*, *question-answer*, and call-out *blocks*, that incorporate 306 enforceable individual policies. Although we can fine-tune a model directly on the raw text, and this may teach the model the vocabulary and general patterns in the text, there is not enough signal for a model to learn to reason about the BCG policies themselves. For clarity in exposition, we showcase only alignment over a base instruction-following model and exclude other bells and whistles that would normally be part of a robust system. Importantly, the internal chatbot application we are imagining is not meant only as an interface to retrieve facts or knowledge about the BCGs but as a general-purpose question-answering service that uses BCG policies as constraints to its responses about various topics.

FRAMERS

The Framers module identifies the knowledge that users consider essential to the application (or domain) so that it can be codified for customization of the LLM model and validation of its results. In a word, Framers *frames* the problem space so that it can be leveraged down the line by the rest of the system. In our running example, this means leveraging the structure and content of the IBM BCGs to create fine-tuning data suitable for model alignment. As mentioned earlier, directly fine-tuning a language model with policy documents would endow it with policy-related vocabulary but would not give it the ability to respond with contextually relevant policy information or to assess policy compliance.

Hence, we proceed to create *instruction*-style data[9,10,] which consists of examples of policy-relevant instructions for various tasks, and *scenario*-style data (discussed below) to align models to the type of tasks that users will need, including identifying the relevant policies for a given situation or whether or not a scenario is compliant. Manually creating sufficient training data is expensive, hence, we adopt a hybrid approach where we create some *seed* data in both styles and use LLMs to create synthetic data to augment this dataset. Both datasets require extraction of paragraphs and self-sufficient atomic policies from the BCG document. We could also use other sources of data, such as policy training materials that contain questions and answers related to policies.

For the first style of data, we extract three types of seed data: 1) *topic-paragraph*, corresponding to topics and paragraphs in the document; 2) *question-answer*, corresponding to the question and answers provided in the document; and 3) *blocks*, corresponding to call-out blocks that highlight a policy scenario. These correspond to only two different task instructions: *summarization* and *question-answering*. The small quantity of seed data we have and these two instruction types alone do not enable the model to generalize. Therefore, we prompt another LLM to generate *synthetic* data based on these seed data. We find that

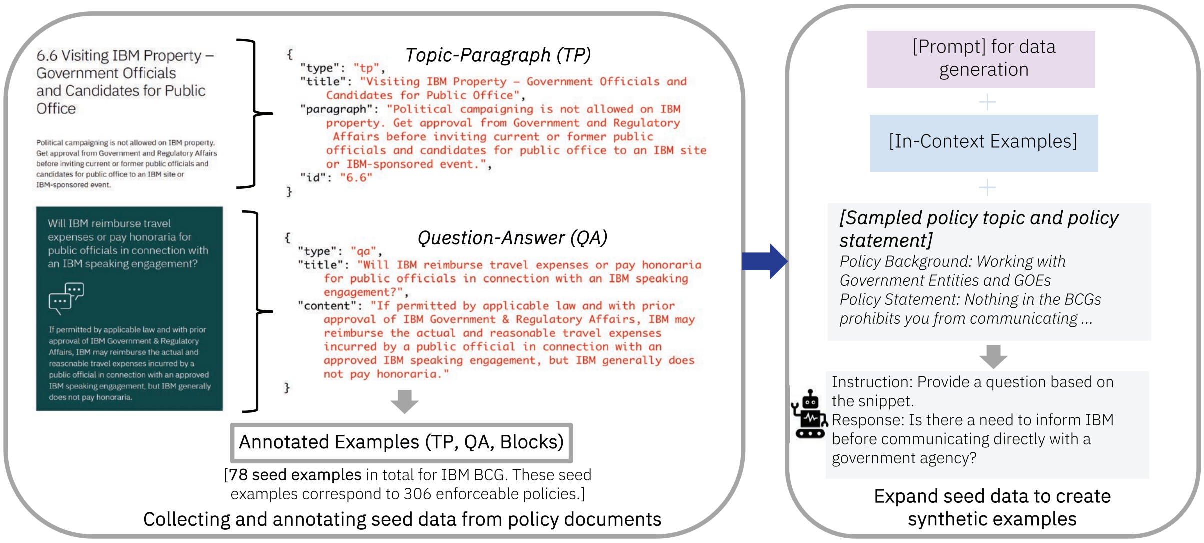
[c] https://www.ibm.com/investor/governance/business-conduct-guidelines

LLMs are adept at producing a diverse variety of task instructions, even starting from just two seed tasks. This is true for powerful LLMs such as *LLama2-70B*, *Falcon-180B*, and *Mixtral-8* × *7B*, which use just a few in-context examples. We create 100,000 synthetic examples and are left with roughly 76,000 examples to train the model after filtering malformed examples and withholding a small fraction as validation/test data. A depiction of creating instruction style data is provided in Figure 3.

For the scenario-style data, we start by manually creating real-world situations that comply, violate, or are ambiguous (they require extra information to make a decision) for a small number of policies in the IBM BCGs. For example, for the policy *It is your responsibility to maintain IBM confidential and proprietary information*, a compliant (fictional) scenario is *Asha was asked by her good friend about a deal that IBM was involved in, but she knew it was confidential. So she politely declined to answer.* Noncompliant scenarios are created similarly; contrastive scenarios can be used to align models to policies in a fine-grained way. As manual data creation for scenarios can be expensive, we leverage these *seed* data to create large numbers of synthetic scenarios for every policy in the document by appropriately prompting LLMs.

Although asking LLMs to generate synthetic data usually results in knowledge of good quality and reasonable variety, the quality may be further improved using ontologies with reasoning over relationships. We use structured, factual knowledge contained in open ontologies such as Wikidata and ConceptNet to systematically generate data to cue scenarios with a thorough domain vocabulary. For example, the BCGs contain structural hierarchies within and outside of IBM, such as organizations (like suppliers, competitors, and government entities), departments (for instance, legal, human resources, and accounting), assets (like products, facilities, systems, and intellectual property), and people (for example, IBM employees, government officials, and family members). Further, the BCGs state what different entities are and how they are organized and related: for example, what constitutes a government entity, how an employee relates to a manager, and what kind of information a data processor handles.

We apply this knowledge to construct a domain-specific ontology for the BCGs to complement the entities and relationships that are extracted automatically from Wikidata. We extract the inheritance hierarchies and ancillary entities (e.g., locations) based on the semantics of the relationships. This ontological structure clarifies ambiguous statements (as seen earlier with "an employee may not work for a competitor") by explicitly stating the relationship between employee and organization (for which the employee works), organization and competitor, and whether "works" is a permissible relationship between employee and competitor. The ontological structure is also used to check the coverage of the synthesized dataset as all ontology terms and relationships are expected to appear at least once in the dataset.



6.6 Visiting IBM Property – Government Officials and Candidates for Public Office

Political campaigning is not allowed on IBM property. Get approval from Government and Regulatory Affairs before inviting current or former public officials and candidates for public office to an IBM site or IBM-sponsored event.

```
{
  "type": "tp",
  "title": "Visiting IBM Property – Government Officials and
      Candidates for Public Office",
  "paragraph": "Political campaigning is not allowed on IBM
      property. Get approval from Government and Regulatory
       Affairs before inviting current or former public
      officials and candidates for public office to an IBM site
      or IBM-sponsored event.",
  "id": "6.6"
}
```

Will IBM reimburse travel expenses or pay honoraria for public officials in connection with an IBM speaking engagement?

If permitted by applicable law and with prior approval of IBM Government & Regulatory Affairs, IBM may reimburse the actual and reasonable travel expenses incurred by a public official in connection with an approved IBM speaking engagement, but IBM generally does not pay honoraria.

```
{
  "type": "qa",
  "title": "Will IBM reimburse travel expenses or pay honoraria
      for public officials in connection with an IBM speaking
      engagement?",
  "content": "If permitted by applicable law and with prior
      approval of IBM Government & Regulatory Affairs, IBM may
      reimburse the actual and reasonable travel expenses
      incurred by a public official in connection with an
      approved IBM speaking engagement, but IBM generally does
      not pay honoraria."
}
```

Annotated Examples (TP, QA, Blocks)
[78 seed examples in total for IBM BCG. These seed examples correspond to 306 enforceable policies.]
Collecting and annotating seed data from policy documents
[Prompt] for data generation
+
[In-Context Examples]
+
[Sampled policy topic and policy statement]
Policy Background: Working with Government Entities and GOEs
Policy Statement: Nothing in the BCGs prohibits you from communicating ...

Instruction: Provide a question based on the snippet.
Response: Is there a need to inform IBM before communicating directly with a government agency?

FIGURE 3. (a) Creating instruction style seed data from policy documents and (b) using it to generate synthetic data using LLMs in a few-shot setting.

INSTRUCTORS

Through Instructors, instilling the desired values and behaviors for alignment is performed using instruction data and human guidance. This is achieved by using supervised fine-tuning (SFT) on high-quality demonstrations of desired behavior, and/or reinforcement learning fine-tuning (RLFT) to optimize the rewards that evaluate preferences over LLM behavior. For the BCGs, these algorithms help the LLM follow various implicit values/behaviors, expressed in the document, through instruction and scenario data generated by the Framers component.

Regulatory documents typically reflect multiple, sometimes conflicting, values or desired behaviors that LLMs need to be aligned with, requiring techniques for the aggregation of these values and behaviors. Instructors allows for the training of reward models from both preference data and binary labels. These rewards assess how well the LLM output aligns with each individual value and desired behavior considered under a use case. Instructors allows for the elicitation of the values or principles we want the LLM to follow as well as the relative importance among them to resolve possible conflicts. Then, RLFT is used to align the LLM with these values based on their relative importance. Instructors also allows for inferring the relative importance of each value from the context in which the LLM is used.[7] Finally, in a low-resource regime, fine-tuning requires parameter-efficient optimization strategies such as (quantized) low-rank adaptations.[11]

AUDITORS

The Auditors component ensures that the data from Framers and the methods from Instructors have resulted in a well-performing model with respect to all the desired criteria, including particular contextual regulations. In general, model evaluation can be categorized along three axes:

1) *When*: The moment in the development lifecycle of the model or application when the evaluation is performed (e.g., during training to ensure that the model is capable with respect to general desirable abilities and/or particular regulations; after training, once a model checkpoint is deemed sufficiently performant, to establish whether the model satisfies criteria that are too costly to be checked during training; or after deployment to ensure that no unexpected or unaccounted-for behavior is encountered).
2) *What*: The type of data that are used during the evaluation (e.g., established benchmarks for testing general-purpose abilities, general-purpose alignment data to test for general human preferences, or handcrafted, domain-specific data to ensure adherence to particular desired criteria).
3) *How*: How the auditing or evaluation methodology is performed and by whom (e.g., automated evaluation based on well-defined benchmarks, human-in-the-loop red-teaming of models, or a combination of both).

Systematic evaluation of models for particular contextual regulations requires specially crafted data as general benchmarks that cover specific regulations are unlikely to exist. Hence, a domain-specific evaluation is carried out in two steps. First, the model is evaluated for alignment against a small, curated dataset of test cases. Then, red-teaming[12] is utilized to uncover potential deficiencies in the aligned model. This red-teaming step helps to dynamically extend the datasets that can be used across the lifecycle for subsequent iterations of the aligned model.

Red-Teaming

We find that red-teaming for adherence to particular contextual regulations is particularly effective when comparing the output generated by two models side by side: one aligned model that Instructors has trained with the data from Framers, and one unaligned model that has not seen any of the particular regulation data. Given these two models, red-team members are asked to craft prompts that test for adherence to the regulations or policies of interest. Red-team members grade the responses along different dimensions such as faithfulness and completeness, providing detailed comments on the output quality whenever binary grading does not provide a sufficient signal. The data collected during this red-teaming can be used to develop further insights into improving the aligned model as follows:

- *Both the aligned and unaligned models generate aligned responses*: The corresponding test cases are deemed “straightforward” examples, where the input is not difficult enough such that alignment to the specific regulation is required to provide a relevant and accurate answer.
- *The unaligned model provides better responses than the aligned model*: This situation indicates that conflicting values and/or poor-quality instruction data may have negatively impacted the helpfulness of the aligned model.
- *The unaligned model fails to follow the particular regulations being tested, while the aligned model does*: These are great examples! They demonstrate

the need for specific alignment to the particular policies to accurately respond to the query.

- *Both the aligned and unaligned models fail to follow the particular regulations being tested*: These are the "hardest" examples, indicating that more or better data are required for the aligned model to learn the particular regulation.

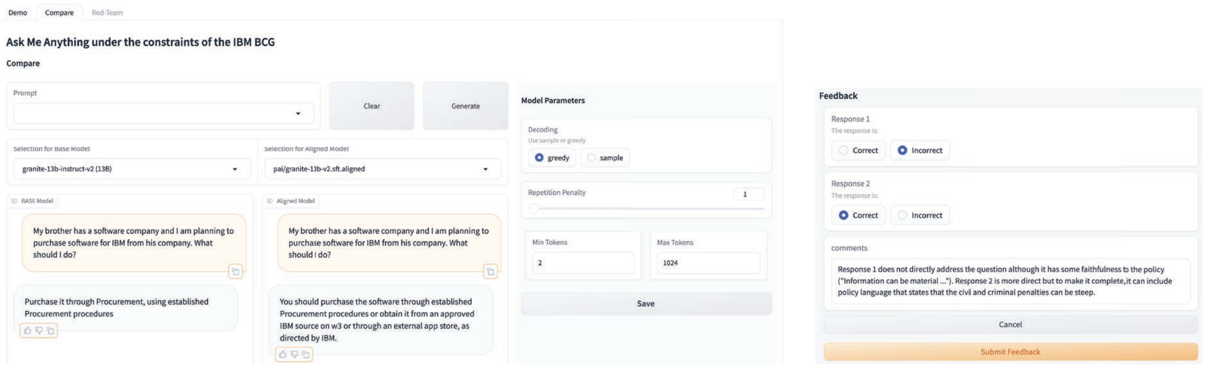
Recall that the main priority of the Auditors is to validate initial alignment and establish baselines. Baselines are then used to continually control and ensure the desired behavior of the LLM model. Indeed, the key aspect of Auditors is that evaluation *never* gets completed: it is a dynamic activity that lasts as long as the deployed model or application is user facing.

SUMMARY DEMONSTRATION

We demonstrate Alignment Studio by aligning an IBM Granite model[13] to the IBM BCGs. We create the seed *instruction* style data using the methodology described in the Framers section and use SFT to align the model. We developed a user interface (UI) to achieve three goals: 1) prompt the aligned model to obtain responses, 2) compare the responses of the aligned and unaligned models, and 3) capture user preference between the responses of the unaligned and aligned models, along with the capability to perform red-teaming. The UI, illustrated in Figure 4(a), contains three tabs: one corresponding to each of the aforementioned goals. Using the UI, we illustrate an example that compares the responses of the unaligned and aligned models to a user prompt about purchasing software. The aligned model produces a response that is more faithful to the policy.

As another example, consider the following prompt: *"I want to create a plan that will help me allocate time between my work at IBM and volunteering as an open source developer. What should I be careful about?"* The unaligned model produces the response *"You should be careful about spending too much time on open source projects as this can affect your work performance at IBM"*, while the aligned model responds, *"Consult IBM's Open Source Participation Guidelines and follow them."* The unaligned model response, although reasonable, is neither helpful nor faithful. The aligned model clearly and concisely states the guidelines that need to be consulted.

Finally, consider the prompt *"My brother has a generative AI start-up company. What kinds of insider information about IBM's new venture fund can I send him?* The response of the unaligned model is, *"Information that is not available to the general public that could influence a reasonable investor to buy, sell or hold a company's securities is considered inside information. Information can be material even if it would have only a small impact on the price of the securities.* The aligned model responds, *"None. Insider information about IBM's financial performance, acquisitions, divestitures, or other significant activities, or inside information about IBM or any other company, is strictly confidential and should not be disclosed to anyone outside of IBM, including family members or IBM colleagues who do not have a business need to know. Additionally, buying or selling securities of a client, IBM Business Partner, or other party based on inside information is unethical and illegal. It is important to consult IBM Counsel if you have any questions or concerns about what you can or cannot do with inside information."* The unaligned



Demo Compare Red-Team

Ask Me Anything under the constraints of the IBM BCG

Compare

Prompt

Clear Generate

Selection for Base Model: granite-13b-instruct-v2 (13B)

Selection for Aligned Model: pai/granite-13b-v2.sft.aligned

BASE Model

My brother has a software company and I am planning to purchase software for IBM from his company. What should I do?

Purchase it through Procurement, using established Procurement procedures

Aligned Model

My brother has a software company and I am planning to purchase software for IBM from his company. What should I do?

You should purchase the software through established Procurement procedures or obtain it from an approved IBM source on w3 or through an external app store, as directed by IBM.

Model Parameters

Decoding
Use sample or greedy
● greedy ○ sample

Repetition Penalty: 1

Min Tokens: 2

Max Tokens: 1024

Save

(a)

Feedback

Response 1
The response is:
○ Correct ● Incorrect

Response 2
The response is:
● Correct ○ Incorrect

comments

Response 1 does not directly address the question although it has some faithfulness to the policy ("Information can be material ..."). Response 2 is more direct but to make it complete,it can include policy language that states that the civil and criminal penalties can be steep.

Cancel

Submit Feedback

(b)

FIGURE 4. UI for (a) comparing the responses of the unaligned and aligned model for a given prompt and (b) evaluating responses for correctness and providing feedback on their quality.

model is incorrect as it does not directly address the question. The aligned model's response, although correct, can be improved to include more policy details. This additional information, given through the feedback UI shown in Figure 4(b), can be used by developers to improve the aligned model.

Furthermore, we used the UI to conduct a masked "taste test" between the unaligned and aligned models. IBM employees who are annually tested on the BCGs served as the participants. Among 10 evaluation prompts and 36 participants, we found that the responses from the aligned model were significantly preferred to be more governed by the BCGs (83.9%; p value 1.2e-15 using a paired t-test).

CONCLUSION

We presented a principled approach to align LLMs to particular contextual regulations, along with a robust and extensible architecture for achieving this. Our methodology is flexible and we demonstrated it by aligning an LLM to the IBM BCGs. Future work includes expanding to other unstructured forms of value specifications and adding semiautomated approaches to eliciting misaligned responses.[14]

ACKNOWLEDGMENTS

The authors thank Payel Das, Michael Hind, Miao Liu, Ronny Luss, Jacquelyn Martino, Erik Miehling, John T. Richards, Matthew Riemer, Skyler Speakman, John Wamburu, and Dennis Wei for comments and discussions.

REFERENCES

1. L. Lessig, "The new Chicago school," *J. Legal Stud.*, vol. 27, no. S2, pp. 661–691, Jun. 1998, doi: 10.1086/468039.
2. Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," 2022, *arXiv:2212.08073*.
3. H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale, "Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback," 2023, *arXiv:2303.05453*.
4. T. Sorensen et al., "A roadmap to pluralistic alignment," in *Proc. Int. Conf. on Mach. Learn.*, Vienna, Austria, 2024, pp. 46,280–46,302.
5. K. R. Varshney, "Decolonial AI alignment: Openness, viśeṣa-dharma, and including excluded knowledges," in *Proc. AAAI Conf. AI Ethics Soc.*, 2024, pp. 1467–1481.
6. Z. Sun et al., "Principle-driven self-alignment of language models from scratch with minimal human supervision," in *Proc. 37th Conf. Neur. Inf. Process. Syst.*, 2024, pp. 2511–2565.
7. I. Padhi et al., "Comvas: Contextual moral values alignment system," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 8759–8762.
8. S. Lazar, "Frontier AI ethics," *Aeon*, Feb. 13, 2024. [Online]. Available: https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai
9. Y. Wang et al., "Self-instruct: Aligning language models with self-generated instructions," in *Proc. Annu. Meeting Assoc. Comput. Ling.*, Jul. 2023, pp. 13,484–13,508.
10. O. Honovich et al., "Unnatural instructions: Tuning language models with (almost) no human labor," in *Proc. Annu. Meeting Assoc. Comput. Ling.*, Jul. 2023, pp. 14,409–14,428.
11. E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
12. D. Ganguli et al., "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," 2022, *arXiv:2209.07858*.
13. IBM Research, "Granite foundation models." Accessed: Sep. 1, 2024. [Online]. Available: https://www.ibm.com/downloads/cas/X9W4O6BM
14. G. Kour et al., "Unveiling safety vulnerabilities of large language models," in *Proc. Workshop Natural Lang. Gener. Eval. Metrics (GEM)*, 2023, pp. 111–127.

SWAPNAJA ACHINTALWAR is a software engineer at IBM India Private Limited, Pune, Maharashtra 411057, India. Her research interests include full-stack development with a focus on trustworthy artificial intelligence. Achintalwar received her bachelor's degree in information technology from Maharasthra Institute of Technology. Contact her at swapnaja.achintalwar@ibm.com.

IOANA BALDINI is a senior research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. Her research interests include computer architecture and runtime systems, cloud infrastructure, and applied natural language processing, and her current focus is on responsible AI. Baldini received her Ph.D. degree from the University of Toronto. Contact her at ioana@us.ibm.com.

DJALLEL BOUNEFFOUF is a senior research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. His research interests include reinforcement learning, active learning, and natural language processing. Bouneffouf received his Ph.D. degree in computer science from University of Paris-Saclay. Contact him at djallel.bouneffouf@ibm.com.

JOAN BYAMUGISHA is a staff research scientist at IBM Research Africa, Johannesburg, 2001, South Africa. Her research interests include knowledge-based biomedical text processing, Bantu language text generation, and linguistic computational formalisms to augment corpus-based natural language processing. Byamugisha received her Ph.D. degree in computer science from the University of Cape Town. Contact her at joan.byamugisha@ibm.com.

MARIA CHANG is a senior research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. Her research interests include neurosymbolic methods for natural language understanding, knowledge representation and reasoning, and trustworthy AI. Chang received her Ph.D. degree in computer science from Northwestern University. She is an elected member of the Association for the Advancement of Artificial Intelligence Executive Council. Contact her at maria.chang@ibm.com.

PIERRE DOGNIN is a research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. His research interests include trustworthy artificial intelligence and machine learning, generative models for natural language, multimodal statistical modeling, and graphical models. Dognin received his Ph.D. degree in electrical engineering from the University of Pittsburgh. Contact him at pdognin@us.ibm.com.

EITAN FARCHI is a distinguished engineer at IBM Research, Haifa, 3498825, Israel. His research interests include security of machine learning (ML)-based systems, application of game theory to the analysis of ML systems, and using ML in the nonfunctional testing of microservices. Farchi received his Ph.D. degree in mathematics from Haifa University. Contact him at farchi@il.ibm.com.

NDIVHUWO MAKONDO is a staff research scientist and manager at IBM Research Africa, Johannesburg, 2001, South Africa, and a visiting researcher and research associate at the Robotics Autonomous Intelligence and Learning Lab, School of Computer Science and Applied Mathematics, University of the Witwatersrand. His research interests include knowledge representation and reasoning, natural language processing, and neurosymbolic learning and reasoning. Makondo received his Ph.D. degree in computational intelligence and systems science from the Tokyo Institute of Technology. He is a senior member of the South African Institute of Electrical Engineers. Contact him at ndivhuwo.makondo@ibm.com.

ALEKSANDRA MOJSILOVIĆ is a senior director at Google Research, New York, NY, USA. She was previously with IBM Research, Yorktown Heights, NY, USA, where she led the Foundations of Trustworthy AI department. Her research interests include machine learning, multidimensional signal processing, and data science. Mojsilović received her Ph.D. degree in electrical engineering from the University of Belgrade. She is a Fellow of IEEE. Contact her at sashym@gmail.com.

MANISH NAGIREDDY is a research software engineer at IBM Research, and MIT-IBM Watson AI Lab, Cambridge, MA, 02142, USA. He is currently interested in participatory evaluations of language models and uncertainty quantification for generative tasks. Nagireddy received his B.S. degree in statistics, machine learning, and computer science from Carnegie Mellon University. He is a member of the Association for the Advancement of Artificial Intelligence (AAAI) and Association for Computing Machinery (ACM). Contact him at manish.nagireddy@ibm.com.

KARTHIKEYAN NATESAN RAMAMURTHY is a principal research scientist and manager at IBM Research, Yorktown Heights, NY, 10598, USA. His research interests include trustworthy machine learning, applied algebraic topology, and networked data models. Natesan Ramamurthy received his Ph.D. degree in electrical engineering from Arizona State University. He is a Senior Member of IEEE. Contact him at knatesa@us.ibm.com.

INKIT PADHI is a senior research engineer at IBM Research, Yorktown Heights, NY, 10598 USA. His research interests lie at the intersection of machine learning and natural language processing. Padhi received his M.S. degree in computer science from the University of Southern California. Contact him at inkit.padhi@gmail.com.

ORNA RAZ is a researcher at IBM Research, Haifa, 3498825, Israel. Her research interests lie at the intersection of artificial intelligence and software engineering. Raz received her Ph.D. degree in software engineering and computer science from Carnegie Mellon University. Contact her at ornar@il.ibm.com.

JESUS RIOS is a research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. His research interests include machine learning, statistical learning, and business applications

of artificial intelligence. Rios received his Ph.D. degree in mathematics and computer science from Universidad Rey Juan Carlos, Spain. Contact him at jriosal@us.ibm.com.

PRASANNA SATTIGERI is a principal research scientist at IBM Research, and MIT-IBM Watson AI Lab, Cambridge, MA, 02142, USA. His research interests encompass include generative modeling, uncertainty quantification, and learning with limited data. Sattigeri received his Ph.D. degree in electrical engineering from Arizona State University. He is a Senior Member of IEEE. Contact him at psattig@us.ibm.com.

MONINDER SINGH is a senior research scientist at IBM Research, Yorktown Heights, NY, 10598, USA. His research interests include artificial intelligence (AI), especially trustworthy AI, machine learning and data mining, and text mining. Singh received his Ph.D. degree in computer and information science from the University of Pennsylvania. Contact him at moninder@us.ibm.com.

SIPHIWE A. THWALA is a research scientist at IBM Research Africa, Johannesburg, 2001, South Africa. His research interests include the extraction and modeling of semantic-based insights from unstructured text, the development of interactive data visualizations, and the use of unsupervised learning for identifying signals from plasma produced by nonthermal diffuse emission in extended radio sources. Thwala has a background in computational astrophysics from the University of the Witwatersrand. Contact him at siphiwe.thwala@ibm.com.

ROSARIO A. UCEDA-SOSA is a senior technical staff member at IBM Research, Yorktown Heights, NY, 10598, USA. Her research interests include knowledge representation and reasoning, ontologies and Semantic Web. Uceda-Sosa received her Ph.D. degree in computer science and engineering from the University of Michigan. She is a senior member of the Association for Computing Machinery. Contact her at rosariou@us.ibm.com.

KUSH R. VARSHNEY is an IBM Fellow at IBM Research, Yorktown Heights, NY, 10598, USA. Varshney received his Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology. He is a Fellow of IEEE. Contact him at krvarshn@us.ibm.com.