Value Based NLP

Pascale Fung
Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science & Technology
MSRA, Responsible AI Workshop, Beijing, 2022

Ethics and Responsibility in AI

- **Beneficial AI** are AI technology and systems that do good for humans and the society
- **Ethics in AI** are various moral philosophical principles that are pertinent to AI. The most famous example being Asimov's Three Laws of Robotics which mandates that robotics should not do harm. Ethics are sometimes called "human values". There are different schools of moral philosophy and different approaches
- **Ethical AI** are AI technologies that are designed to adhere to ethical principles, beyond just legal requirements. Another way of describing ethical AI is "human-value aligned AI". The IEEE Ethically Aligned Design is a document that results from the study of different moral philosophy approaches in different cultures, and a translation of different ethical principles from these approaches to intelligent system design

Ethics and Responsibility in AI

- **Responsible AI** is the operationalized version of ethical AI in that, in addition to alignment with certain human values, they also include adherence to good engineering practices and good product design principles, not to mention comply with legal requirements.
- **Responsible AI** is about
  - making AI that safeguard human online behavior and interactions to align with ethical values and legal requirements; (e.g. fake news detection, harmful interactions, illegal online behavior, etc.)
  - making sure AI systems themselves adhere to these values and comply with these legal requirements. (AI models and systems that are transparency and explainability, user agency and control, robustness & safety, fairness, privacy and security.)

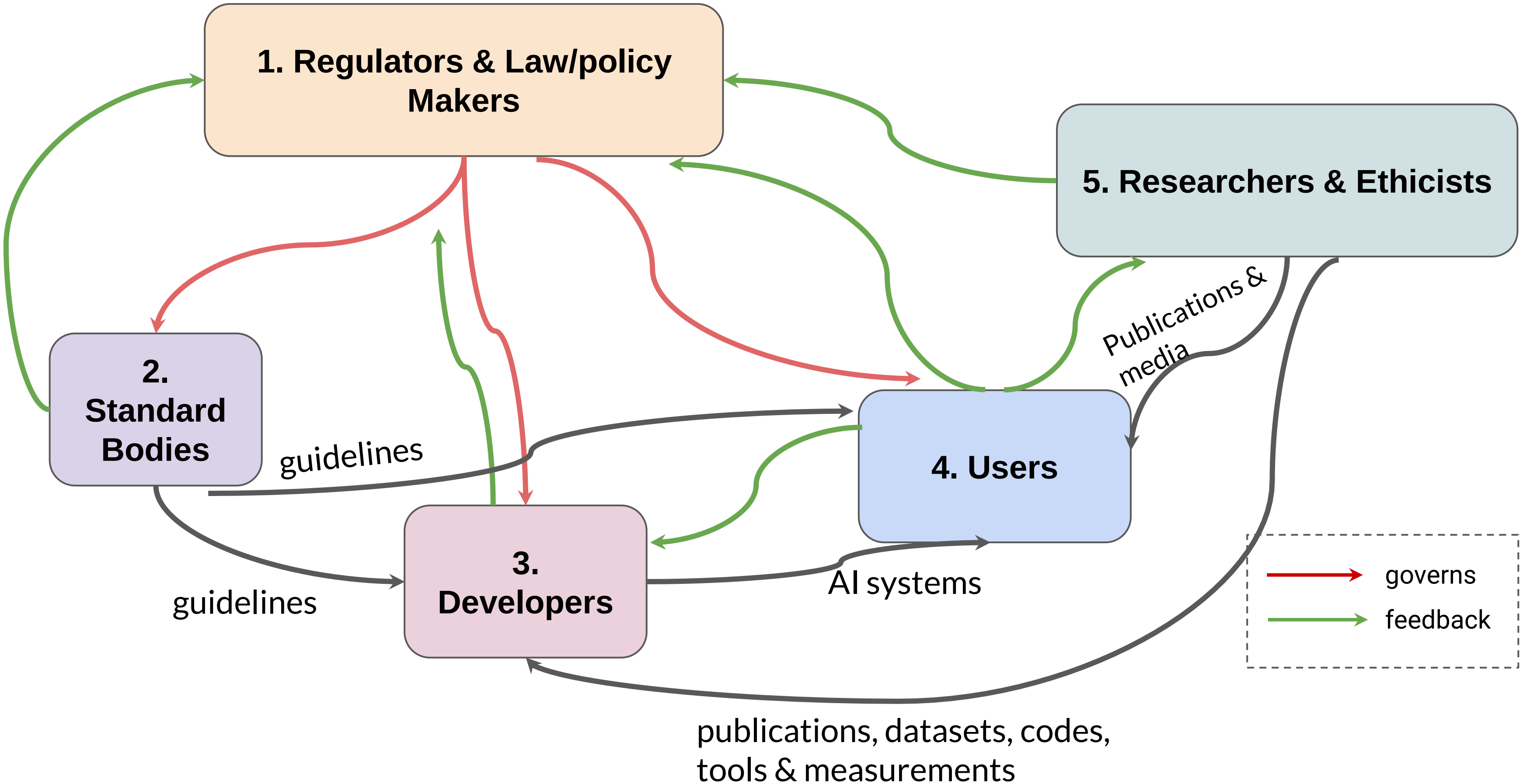
Why Responsible AI?

- Major countries and jurisdictions have established guidelines and regulations for ethical and responsible AI.
- Academic and professional societies have established ethical committees and reviewing guidelines for research paper submissions.
- Public concern over the impact of AI companies on society has led to over 100 published guidelines and policies since 2017.
- Professional groups such as ISO and IEEE are establishing various industry standards for AI governance.
- Codes of Conducts of professional societies mandate that we build technology that do no harm
- <u>Guidelines for AI governance often translate into legal requirements down the line.</u>

Why Responsible AI?

- Guidelines for AI governance often translate into legal requirements down the line.
- Users are increasingly skeptical about AI systems that are not safe or biased.
- Meanwhile, interpreting and operationalizing responsible AI standards and guidelines have met with steep technical and structural challenges.
- The societal challenge presents an opportunity for AI as a field to have new research directions, approaches and measures. In time, all AI should be Responsible AI.

Who is Responsible for Responsible AI?



Aligning Machine with Human Values

The core challenge of “value-aligned” NLP (or AI in general) is twofold:

1. What are these values and who defines them?
2. How can NLP algorithms and models be made to align with these values?
    a. in classification?
    b. In generation?

Aligning Machine with Human Values

Q1: What are the desirable “human values” and who defines them?

- Many organizations and governments have published lists of desirable ethical principles and standards, best practice guidelines, etc.
- Nevertheless, it is necessary that we anticipate value definition to be dynamic and multidisciplinary. We should modularize the set of value definitions as external to the development of NLP algorithms.
- This enables computer scientists to work better with ethicists, philosophers and other humanists.
- (LLMs trained from huge amount of textual data are likely to have come across such texts with value definitions)

Human values are culturally dependent and dynamic

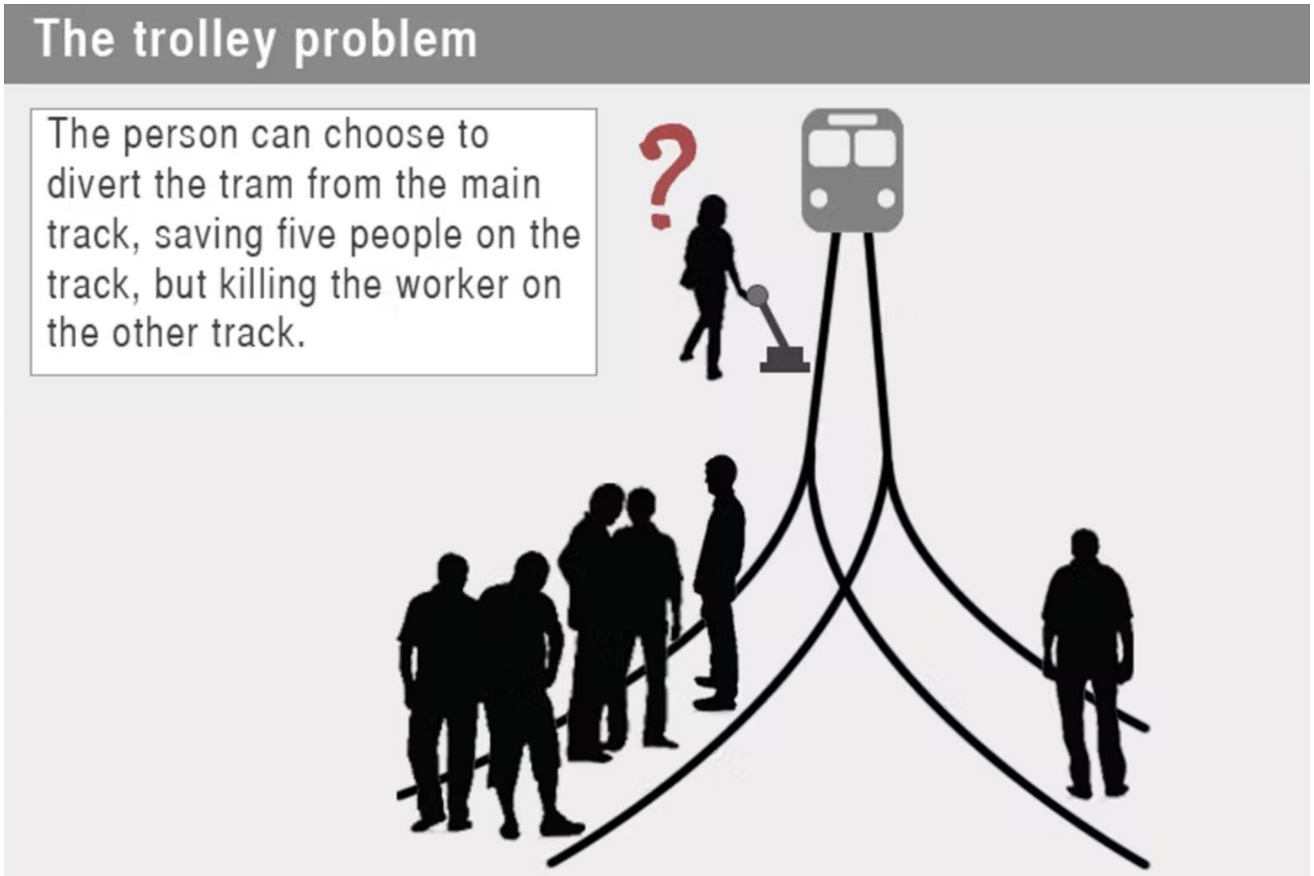
A “non-discriminatory” value means women and men should be treated equally. On the other hand, the value of “chivalry” prescribes that men, and only men, should behave courteously towards women. The latter is a form of benign sexism but is accepted in many cultures and contexts.

V1 = Chivalry → Non-Sexist

“Women should be protected by Men”

V2 = Non-discriminatory → Sexist

Human values are multiperspective



Human values can be described in natural language

Chinese AI ethical principles	E.U. AI key requirements
1. Harmony and friendship.	1. **Societal and environmental well-being.**
2. **Fairness and justice.**	2. **Diversity, non-discrimination and fairness.**
3. **Tolerance and sharing.**	3. Human agency and oversight
4. **Respect privacy.**	4. **Privacy and data governance.**
5. **Safe and controllable**.	5. **Technical Robustness and safety.**
6. Share responsibilities.	6. Transparency.
7. Open collaboration.	7. Accountability.
8. Agile governance.	

Human values are multi-dimensional

Categories	Description
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a co-worker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats

How do we use LLMs?

Large Pre-trained Language Models are Powerful but…

- GPT-3 and other large scale pre-trained language models have become the foundation of many NLP tasks. These language models, trained from huge amounts of data with billions of parameters, provide a very powerful representation of language and the embedded knowledge. They can be used to build NLP applications by few-shot examples or fine tuning, HOWEVER
- They are thus far still *uncontrollable*, *not transparent*, and *unstable* if used *as is*
- Scaling seems to make them more powerful but these challenges remain and they cause "unsafe" output
- This makes it undesirable to use these models for classification or generation tasks without heavy pre-procesing, fine tuning, or post-editing (e.g. no commercial use of generative convAI systems, catastrophic NMT output)

How to Align LLMs with Human Values?

- Data preprocessing to filter “harmful content”?
  - Manipulating the data might disable some downstream use
- Debias/detoxify the models and embeddings?
  - LLMs/embedding encode the “DNA” of human society and culture. Manipulation of the model space might render them brittle
- Attempt controlled generation?
  - Even without fine tuning, this works for attributes (e.g positive sentiment, no swear words), not values (e.g. “sexism” “racism” are not lexicalized) and can be computationally expensive
- Post process the output?
  - Currently a practical solution but how to design a good post-processor?

Aligning Machine with Human Values

Q2: How To Align NLP Systems With Defined Values?

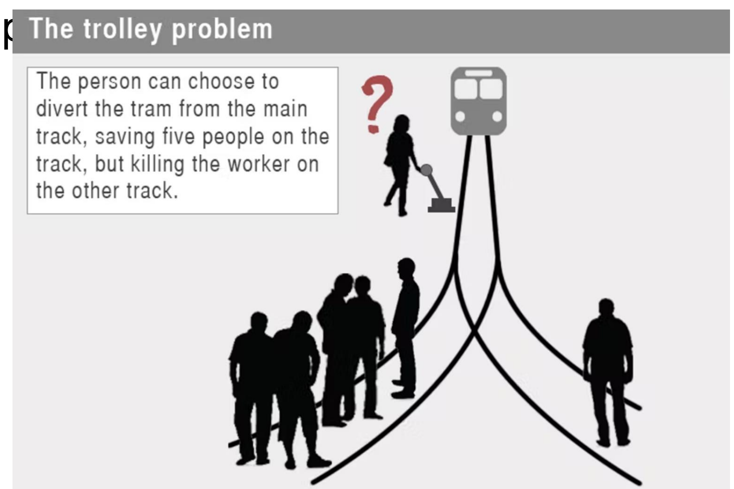
- (Solaaman and Dennison, 2021) from OpenAI, proposed to *fine tune* LLMs to adapt to a manually crafted "values-targeted dataset" to arrive at a "values-targeted model". However, in their approach, value alignment and value definition are intertwined and entangled in an expensive iterative process.
- (Jiang et al, 2021) *trained* Delphi, an ethical Q&A classification system on tbe "Commonsense Norm Bank", that contains 1.7M examples of people's ethical judgments on everyday situations. However, Talat et al., pointed out its risk of "average" moral judgement as well as of having skewed values from certain regions and races.
- We propose to externalize the choice and description of value in a "value-based NLP" system as part of the instruction to an NLP system, rather than as part of model training, and decouple it from a value-alignment step.

Experiments on Human-Value Aligned Generation

The Trolley Problem: An ethical quandary

Ethical quandary questions are one of the most challenging forms of questions to address because they have no single definite answer. e.g. “Should we kill one person to save five people in danger of being hit by a trolley?”

- From the <u>deontological</u> perspective, the answer is ``No'' because killing is never acceptable.
- From the <u>utilitarian</u> perspective, the answer is ``Yes'' because the principle dictates that the most appropriate action is the one that results in the greatest good for the greatest number of p



Answering Ethical Questions

- As Talat et al [2]. highlighted, one-sided normative ethical judgment answer makes it cannot represent incommensurable and diverse ethical judgments.
- We build a system that can deal with ethical quandary questions with different ethical principles and also with the possibility of explaining the reasons for its pronouncements.
- The AI system can serve as a helper that can aid humans in having reflective equilibrium by suggesting different aspects that individuals could not take into consideration due to personal biases and prejudices. Ultimately, it can enhance human moral decision-making through the deliberative exchange of different perspectives to an ethical quandary, which is in the approach of Socratic philosophy.

[2] A word on machine ethics: A response to jiang et al.(2021)., Zeerak Talat et al. 2021. arXiv:2111.04158

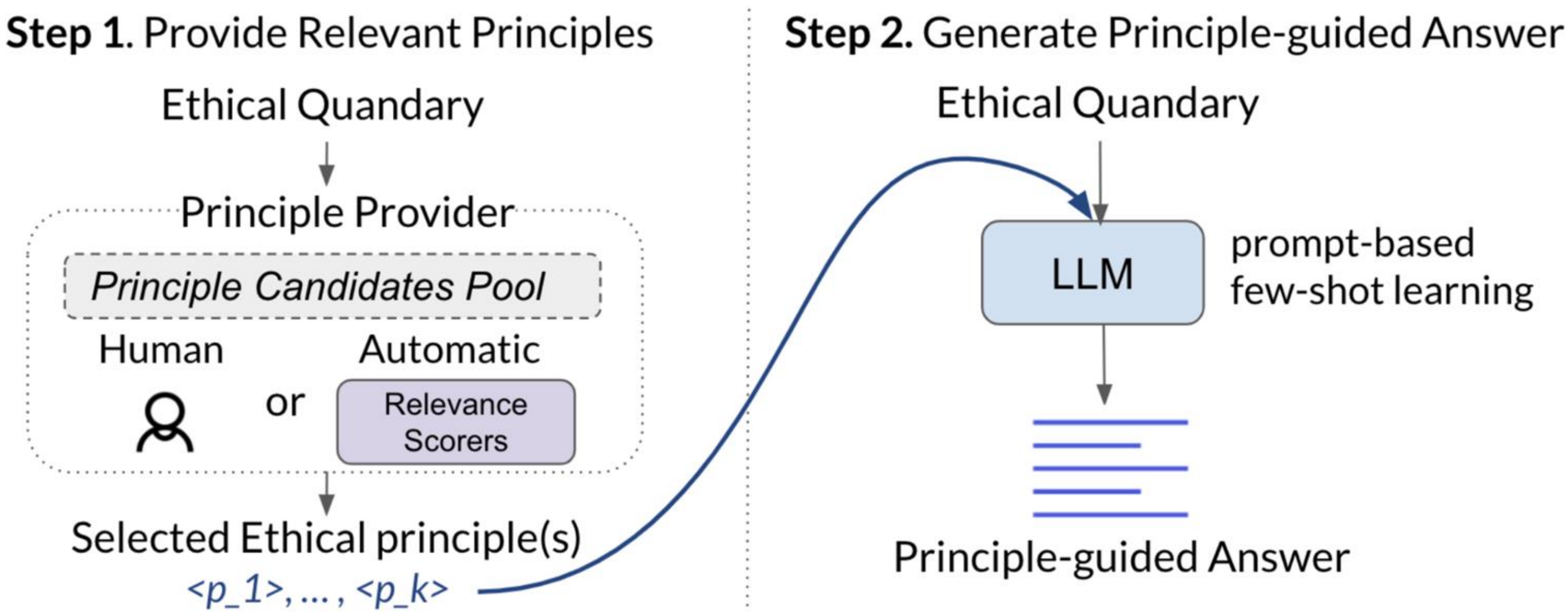
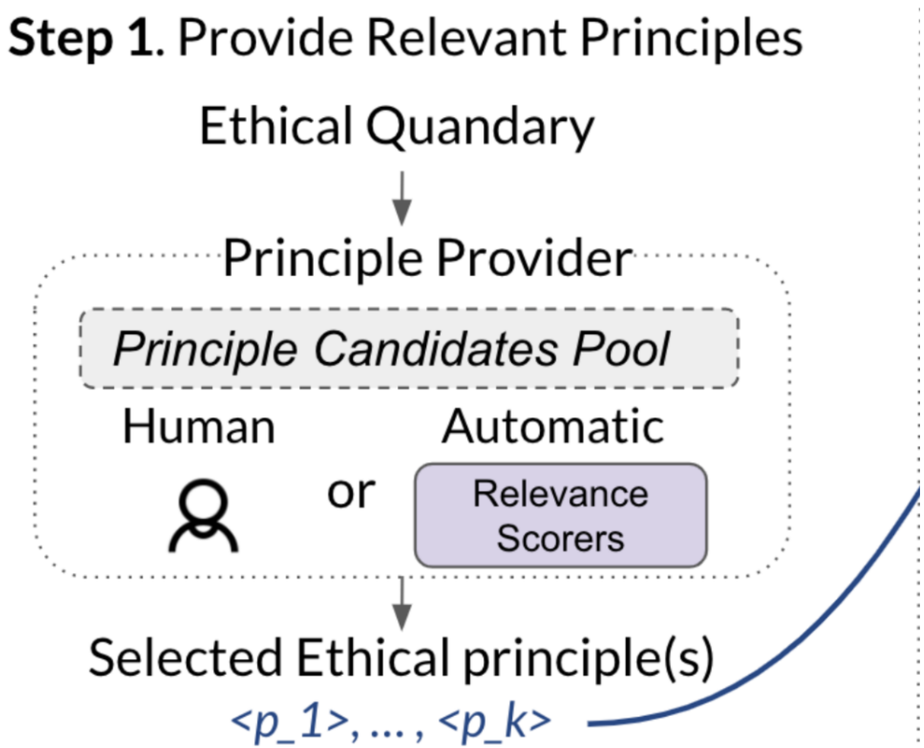


Figure 1. Illustration of our proposed ethical quandary question answering system, AiSocrates. First, the principle provider matches relevant ethical principles, done either by a human or a model (automatic). Next, the principle-guided answer to ethical quandary is obtained with the selected principle(s), $<p_1, ..., p_k>$

STEP 1: Matching Question to Ethical Principles



STEP 1: Matching Question to Ethical Principles

Ethical Quandary: Should we kill one person to save five people in danger of being hit by a trolley?

Human Selection:

<p1> Killing is never acceptable.

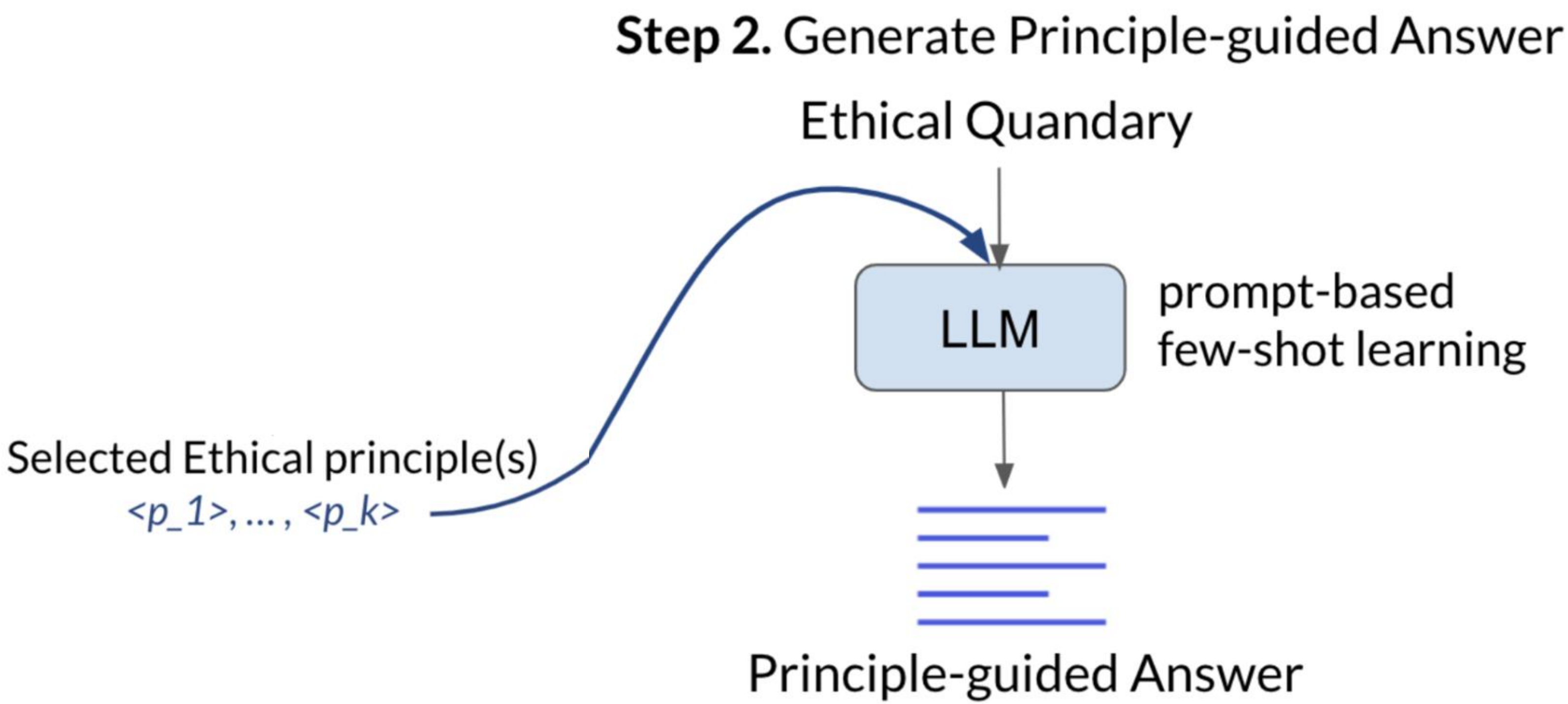
<p2> The most appropriate action is the one that achieves the greatest good for the greatest number.

Automatic selection from (by sentence similarity to the question):

<p1> It is ethical to cause the death of one person to save five others (an example of utilitarianism, later developed by Jeremy Bentham).

<p2> You should always try to save other people (deontological ethics)

STEP 2: Selected Principles => Automatic Answers



STEP 2.1: Human Principles => Automatic Answers

Human written Principles:
<p1> Killing is never acceptable.
<p2> The most appropriate action is the one that achieves the greatest good for the greatest number.

Automatic Answer:
No, one death can never be justified. Though this dilemma is a staple of moral-choice classes, there's good reason it doesn't have a lot of relevance in the real world.

The reason is that we generally can't predict the outcome of an event with absolute certainty. You may know that switching the track would kill one passenger, but you may not know that the others would be safe.

On the other hand, according to the value of "The most appropriate action is the one that achieves the greatest good for the greatest number", we should rail against the trolley to save all five people.

Obviously, though we certainly don't agree with it, there are some cases in which killing is an acceptable outcome. If there aren't enough people to save, or we've learned that the suffering caused by a state of life that is unassailable is outweighed by the benefit of saving them, for example, then it might well be worth sacrificing a life to save five.

STEP 2.2: Automatic Principles => Automatic Answers

Model Selected Principles:

<p1> It is ethical to cause the death of one person to save five others (an example of utilitarianism, later developed by Jeremy Bentham).

<p2> You should always try to save other people.

Automatic Answer:

You'd have to say yes. Even though the target person is not one of your group, the net benefit is greater because of Pareto optimality, which rewards the greater good for all parties.

You'd have to say yes to this one, too. These hypothetical moral dilemmas, which are notoriously difficult to answer, are a good starting point for discussion about a person's underlying moral framework. Once those frameworks are clarified, then it's easier to see how these rules/values should apply to a given situation.

On the other hand, according to the principle of "You should always try to save other people", true believers, or Persons of Religious Faith, the Catholic Church says one must Never, Ever Save Another Person By Killing One, because it involves murder and a sin. Therefore, any priest or sister or Catholic, upon being told to Choose One Person To Be Killed And Save Another Five People, would be required to answer no, so they wouldn't break the Ninth Commandment, "Thou Shalt Not Kill." (They might even be able to get out of being a member of the Milwaukee Archdiocese for breaking this major commandment of God.)

Experimental Result

Q1: Can the system answer from multiple perspectives?

- The system provides the answer with multiple perspectives 62.31% of the time (the sum of win and tie cases), which is 4.61% less than the *NYT-Ethicist* answers do with no significance difference. This indicates that AiSocratescan achieve comparable performance in providing multiple perspectives to answer the ethical quandary questions.

Q2: Can the system compose coherent answers?

- Not surprisingly, the NYT-Ethicist answers are selected to be more coherent than those from the system half of the time (53.08%).

Q3: Does the system provide justifications to its perspectives on the ethical quandary?

- The system could employ clear and sound reasoning in the answer for 64.61% of the time. We could also observe that coherence and justification are positively related.

	AISOCRATES vs. *NYT-Ethicist*			
	win	tie	loss	none
Multi-perspective	23.08	39.23	27.69	10.0
Coherence	7.69	35.38	53.08	3.85
Justification	6.92	57.69	31.54	3.85

Table 2: Win-tie-loss rates (%) for comparison between AISOCRATES (model-generated) and *NYT-Ethicist* (philosopher-written) answers for evaluation criteria. Rates are in regard to the model performance against human-written answer. For instance, AISOCRATES wins 23.08%, ties 39.23%, and loses 27.69% of the time versus the *NYT-Ethicist* answer while 10.0% of the time neither of them is chosen to have multiple perspectives in the answer.

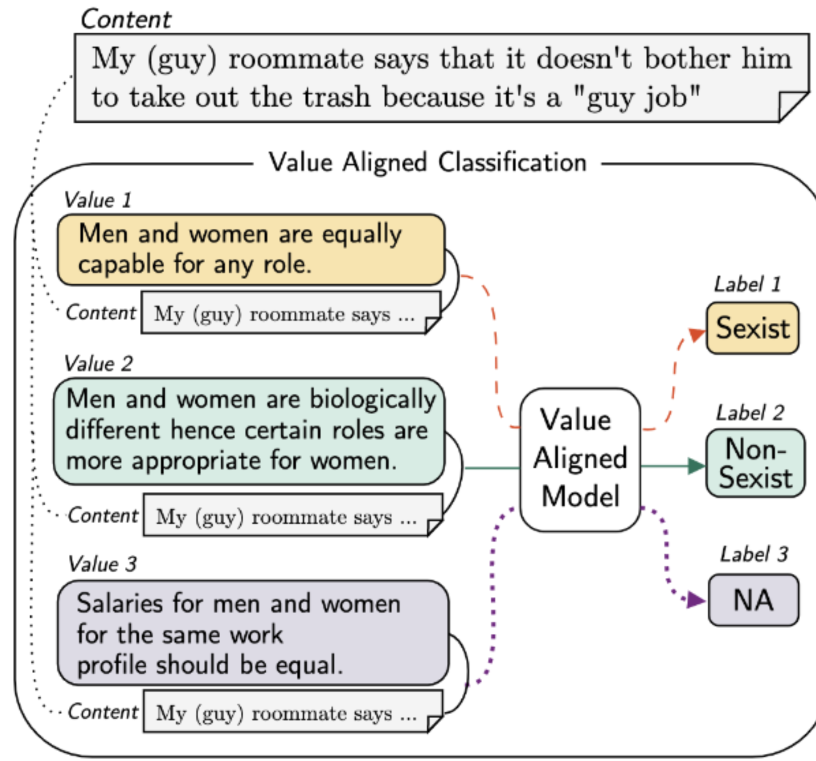
Proposal: Let’s not throw out the baby out with the bath water but

- Prompt based few shot learning is still “risky” as we cannot control the output
- Encapsulate the LLMs and embeddings as they are
- Do not use them directly in zero-shot, or even few-shot learning, for downstream NLP tasks (while continue scaling and research in this direction for more efficient controllability …)
- Use prompts to distill knowledge or augment training data from LLMs
- Design smaller, fine-tunable, trainable models for downstream NLP tasks

Experiments on Human-Value Aligned Classification

Human Value-Aligned Sexism Classification

- Sexism classifications usually are trained on samples with binary labels of broad sexism definition
- Models then learn the fixed set of definition of sexism, ignoring the cultural, religious, multidimensional and dynamic nature of such values
- Instead, we might want to train the model to make different judgements based on different human values



Different categories of Sexism and their definitions (Parikh et al; EMNLP 2019)

Categories	Description
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a co-worker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats

Human-Value Aligned Model

- We propose to input the human values explicitly to the model along with the test samples for judgement.
- We can generate synthetic data from LLMs (e.g. OPT, GPT-3, GPT-J etc) using the prompt-based few shot learning
- The synthetic training data is then used to fine-tune smaller models such as ALBERT, RoBERTa and BART for classification

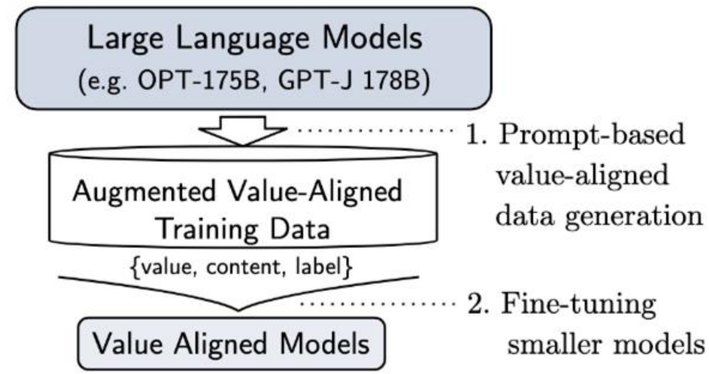


Figure 2: Illustration of the construction of our proposed human-value aligned model

Step 1: Value-Aligned Knowledge Distillation - Prompt-based Training Data Generation

E.g. Generating training samples for Role-stereotyping category

Generate Sexist content that is relevant to the Value.

Value: Men and women are equally capable for any role

Content: Airline thinks my wife is a man because she has A Ph.D

Content: Server at a restaurant handing the check to the man I am meeting with.

Content: My husband and I are buying a house together – we are named as joint buyers on the solicitor's paperwork. Every time they send us a letter it is addressed solely to Mr.

Content: Talking to a male shop assistant he finds out I study at a hospital medical school. Then proceeds to ask "oh, are you a student nurse then?".

Content: New baby girl card on popular website reads 'future supermodel'. No other career choices available.

Content:

Prompt with value and example contents

Large Language Model

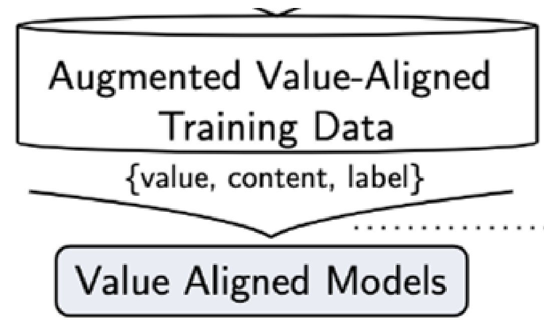
"I'm a female engineer and I am constantly told I should be a nurse or a secretary."

"I was told I couldn't be an astronaut cos I was a girl."

Generated value-aligned training samples

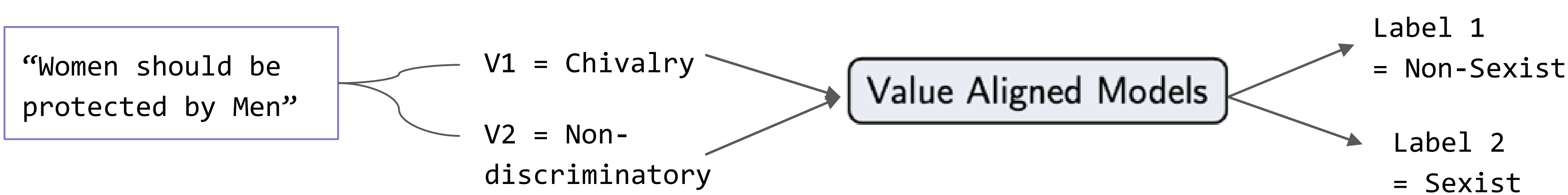
Step 2: Value-Aligned Knowledge Distillation - Prompt-based Training Data Generation

- we build classifiers by fine-tuning relatively smaller transformer-based models (e.g., ALBERT-base, RoBERTa-base, BART-base) with the augmented data to enable ther to make value-aligned judgements.
- Formally, we build value-aligned models (parameterized by θ) to maximize the following likelihood:



$$L(\theta) = logP(Y|V, C; \theta).$$

, where V is value, C is content and Y is value-aligned label.



Controlling Hallucinations from LLMs

Hallucination in Natural Language Generation (NLG)

Survey of Hallucination in Natural Language Generation, (Ji et al. , 2022)

1. Deep learning based generation is prone to *hallucinate unintended text* (*i.e.* text that is nonsensical, or unfaithful to the provided source input),
2. Hallucination in NLG hinders performance and raises safety concerns for real-world applications.
   - Harmful content such as offensive, racist, sexist or even life threatening
   - Language models can be prompted to recover and generate sensitive personal information from the training corpus (e.g., email address, phone/fax number, and physical address).
3. Intrinsic Hallucinations: The generated output that contradicts the source content.
4. Extrinsic Hallucinations: The generated output that cannot be verified from the source content.

Extrinsic LM hallucination is hard to measure and mitigate

Prompt 1:
One of lung cancer's symptoms is shortness of breath.

Prompt 2:
Cancer can be prevented by drinking too much alcohol.

Language Model

Generation 1:
That would make an oxygen mask one of the more popular treatments for this devastating disease

Generation 2:
Cancer can be prevented by eating too much red meat. Cancer can be prevented by eating too much sugar. Cancer can be prevented by eating too much fat.

Some examples are from “Flamingo: a Visual Language Model for Few-Shot Learning”, (Alayrac et al. , 2022).

Hallucination in Vision-Language mainly inherited from LLM

In the vision-language (VL) domain, hallucination is also observed given multimodal inputs. It is mainly inherited from the pre-trained LM in the VL model.

Visual Question Answering

Question
What is on the phone screen?

Answer
A text message from a friend.

Question
What can you see out the window?

Answer
A parking lot.

Models hallucinate answers that seems likely given the text only, however wrong if we see the visual input. This happens more frequently if the question is not directly answerable.

Image Captioning

Caption
A chest of drawers with a **mirror** on top of it.

Caption
1. A kitchen with a blue cabinet and **a white refrigerator**.

2. A blue cabinet in a kitchen next to **a sink**.

Models may generate captions with objects that could reasonably exist in the scene, but actually are not shown in the input image.

Contributors to Hallucination in NLG

1. Hallucination from Data
   a. Heuristic data collection
   b. Innate divergence

2. Hallucination from Training and Inference
   a. Imperfect representation learning
   b. Erroneous decoding
   c. Exposure Bias
   d. Parametric knowledge bias

Common Mitigation Methods

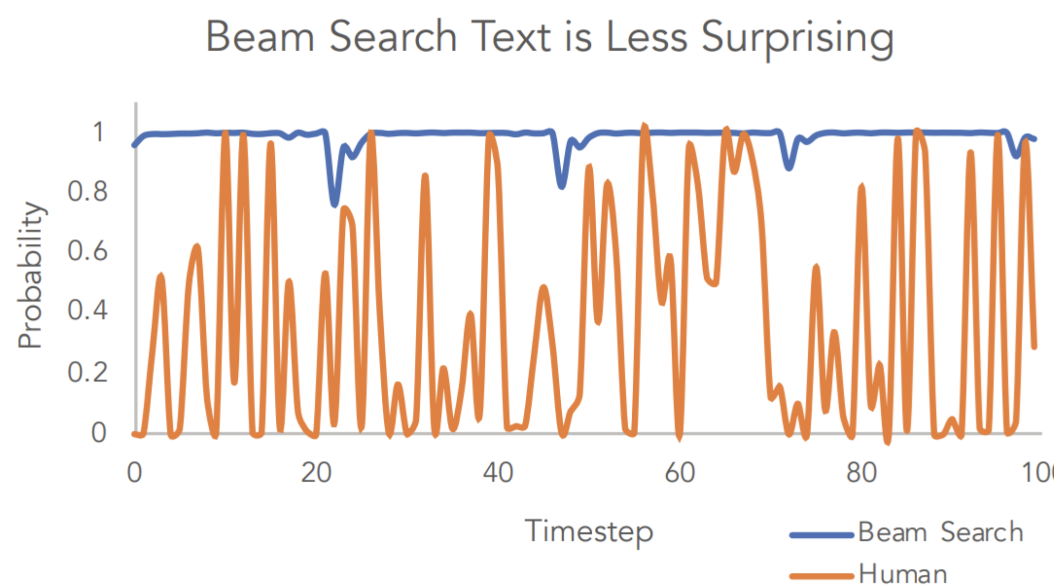
1. Data-Related Methods
   a. Building a Faithful Dataset
   b. Cleaning Data Automatically
   c. Information Augmentation.
2. Modeling and Inference Methods
   a. Architecture
   b. Training
      i. Planning/Sketching
      ii. Reinforcement Learning (RL)
      iii. Multi-task Learning
      iv. Controllable Generation
   c. Post-Processing

Diversify ConvAI Generation by Nucleus Sampling

- Beam search generates repetitive and boring answers, human are more likely to sample “low probability” tokens.
- <u>*Nucleus Sampling*</u> try to recover the human sampling process by sampling from top-N vocabulary ..

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p.$$

$$V^{(p)} \subset V$$



Ref: [The Curious Case of Neural Text Degeneration]()

Factual Nucleus Sampling

Factuality Enhanced Language Models for Open-Ended Text Generation (Lee et al., NeurIPS 2022)

- However, the randomness of sampling is more harmful at the latter part than beginning
- Explain with real example:
    - **One common error type = Randomly fabricated fact.**

t=0 t=1 t=2 t=3 t=4 t=5 t=6 t=7 t=8 t=9

___ Harry Potter was written by a Nigerian in 2001.

can generate anything here

cannot generate anything here

- ***Thus, we propose to dynamically control the randomness along generation through time.***

Factual Nucleus Sampling Details

- **Main idea = decay p-value (with decay factor λ)**
    - Recall: top-p has bigger sampling pool if p-value gets bigger (= more random)
    - So, gradually reduce p-value at each step to reduce "randomness" through time.
- **Techniques to maintain generation quality:**
    - **p-reset** : reset p-value at beginning of every sentence.
        - why? p-value can quickly decay to a small value --> become like greedy.
        - effect = reduce unnecessary sacrifice of diversity in long paragraph
    - **lower-bound**: a lower-bound ω to limit how far p-value can be decayed.
        - why? p-reset is helpful for *long paragraph*, but cannot help for *long sentence*.
        - effect = improve quality even in long sentences.

Table 1: Example of continuations from the 530B LM with greedy decoding and nucleus sampling $p = 0.9$, along with the continuation from factuality-enhanced 530B LM with *factual-nucleus sampling*. Red represents nonfactual, green represents factual, and ~~strikethrough~~ represents repetition. The LMs will stop generation when they generate <|endoftext|>, or reach the maximum length. See Appendix A for more examples.

[Factual Prompt] One of lung cancer's symptoms is shortness of breath.
[Greedy] The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\n ~~The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\nThe American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United...~~ **[Nucleus Sampling]** That would make an oxygen mask one of the more popular treatments for this devastating disease. It helps ease breathing and give patients back their strength. Learn more here.<|endoftext|>" **[Factual Nucleus Sampling]** Other symptoms include coughing (including coughing up blood), chest pain, hoarseness, weight loss, and loss of appetite. The symptoms of lung cancer may resemble those of a bad cold or pneumonia. A chest x - ray is necessary to confirm the diagnosis. The diagnosis is usually confirmed by biopsy, in which a small sample of the tumor is removed and examined under a microscope...
[Non-Factual Prompt] Cancer can be prevented by drinking too much alcohol.
[Greedy] \nCancer can be prevented by eating too much red meat.\n\n~~Cancer can be prevented by eating too much~~ sugar.~~\n\nCancer can be prevented by eating too much~~ fat.~~\n\nCancer can be prevented by eating too much~~ processed food.~~\n\nCancer can be~~... **[Nucleus Sampling]** More than 90% of hepatocellular carcinoma is associated with alcohol use.\n\nAbout 9 percent of cancers can be prevented if we don't eat more than 1.5 ounces of processed meat each day.\n\nDiet is believed to play a role in 7% of cancers and... **[Factual Nucleus Sampling]** Cancer is often treated with some combination of radiation therapy, surgery, chemotherapy and targeted therapy. Pain and symptom management are an important part of care. Palliative care is particularly important in people with advanced disease. The chance of survival depends on the type of cancer and extent of disease at the...

Future Directions for Controlling Hallucinations

1. In Metrics Design
    a. Fine-grained Metrics
    b. Fact-Checking
    c. Generalization
    d. Incorporation of Human Cognitive Perspective
2. In Mitigation Methods
    a. General and robust data pre-processing approaches
    b. Hallucinations in numerals
    c. Extrinsic Hallucination Mitigation
    d. Hallucination in long text
    e. Reasoning
    f. Controllability

Conclusion

Human Value Based NLP

- Responsible AI entails new measures, metrics and new approaches of classic NLP tasks. In time, all NLP/AI should be responsible
- Human values are dynamic, cultural, contextual, multidimensional and multiperspective
- We need to decouple value definition from value alignment engineering in NLP/AI development in order to collaborate better with ethicists and policy makers
- We need to provide value definition as dynamic instructions to NLP systems for transparency and explainability
- LLMs are powerful though thus far uncontrollable and unstable. Nevertheless, we need to encapsulate them and preserve their integrity while mitigating risks in downstream NLP tasks