University of Zurich[UZH]

Institute of Law

The Transformation of International Law: A LLM Design Perspective

UZH, 4.11.25

PD Dr Johan Rochel

Overview

1. An applied research project
2. Three architectural choices
3. The art of taking design decisions

1. An applied research project

Example of a question for the WTO-LLM:

Is the domestic patent norm X (*description of the domestic patent law*) compatible with Article 30 of the TRIPS Agreement (the so-called *three-step test* for exceptions to patent rights)?

Example of a question for the IHL-LLM:

Country X is experiencing high levels of violence in the region where minority A has established a *de facto* administration.

Do the official armed forces of X have the right to attack targets (e.g. energy infrastructure) found in the specific area where minority A lives?

Which legal-philosophical challenges do you face in designing such an LLM?

- Pilot-project of an LLM for international humanitarian law (IHL)
- Design and fine-tune pre-trained open-source model Llama 3.1 (70B) with IHL data (now further test with Apertus)
- General ambition: facilitate access to legal knowledge about IHL, especially in low-resources contexts

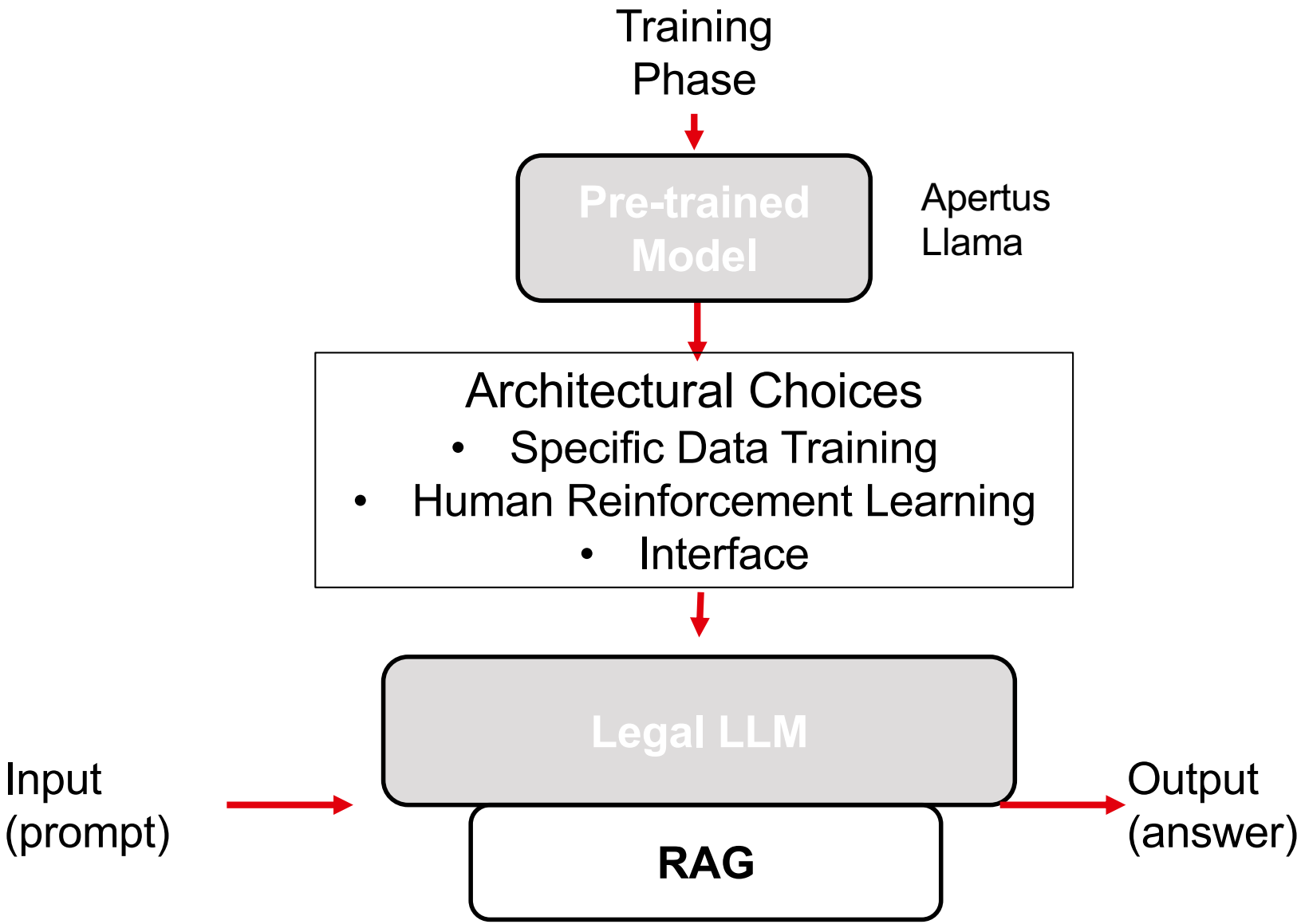
(Devidal 2024; Raymond/Card 2015)

(forthcoming Rochel *Nordic Journal of International Law* 2026)

An LLM is…

- A computational system
- Trained on vast amounts of data
- Fine-tuned by human interventions
- Aiming to become a plausible next-word-predictor
- Generating content hopefully reflective of the world

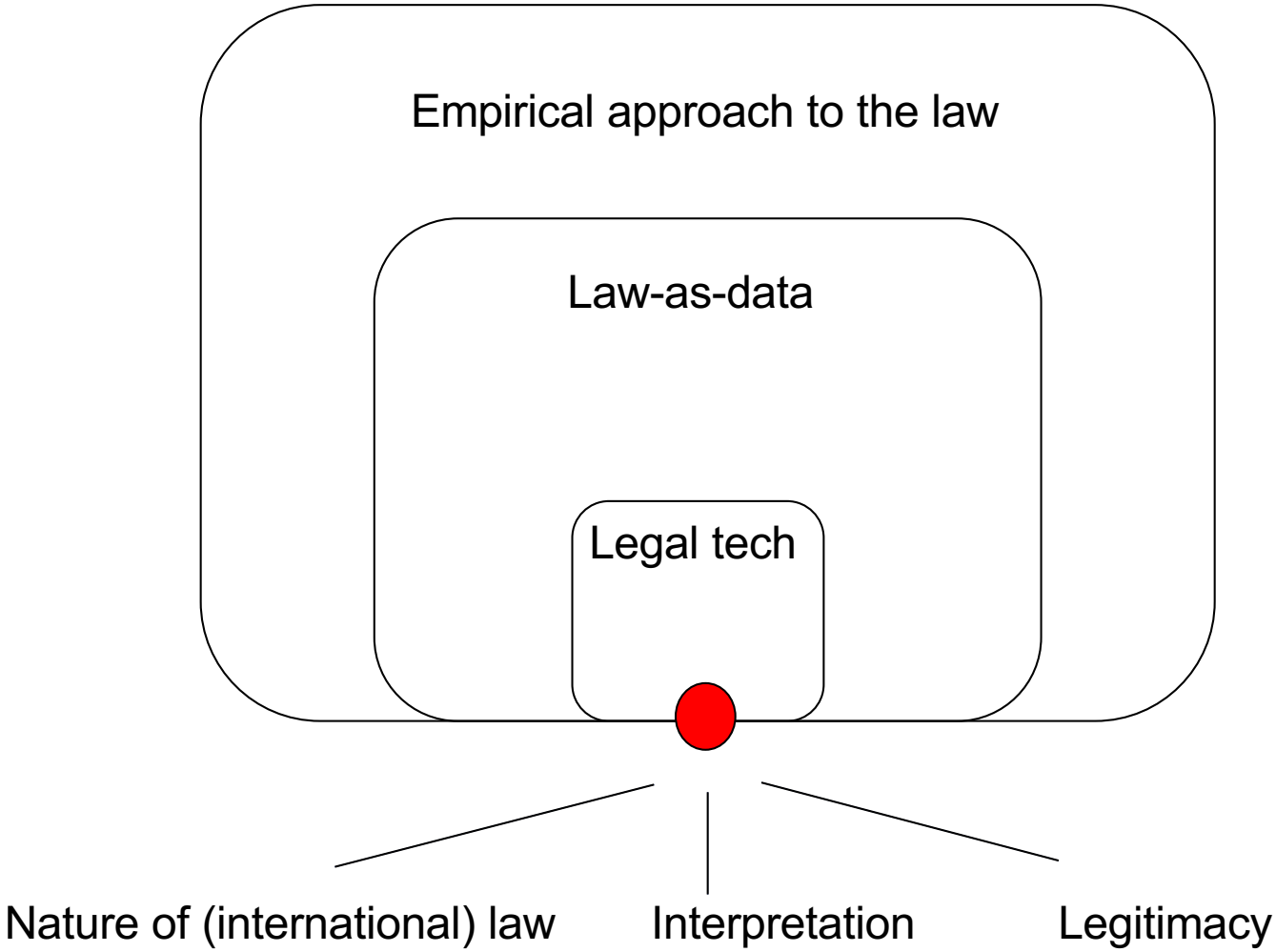
(Dahl et al. 2024, 66)


Training Phase
Pre-trained Model
Apertus Llama
Architectural Choices
• Specific Data Training
• Human Reinforcement Learning
• Interface
Legal LLM
RAG
Input (prompt)
Output (answer)

My goals:

- Unpacking and addressing a set of legal-philosophical questions about (international) law
- Emphasizing importance of thinking about the user and his/her perspective
- Illustrating that stakes are high – design decisions impact international law practices and transform international law

(Deeks 2020)



(Livermore/Rockford 2019;Altwicker 2019;Alschner 2020; Johns 2021;Nelson 2023; Arvidsson/Noll 2023)

Well-identified general challenges of LLM:

- IP issues related to training data, output, model

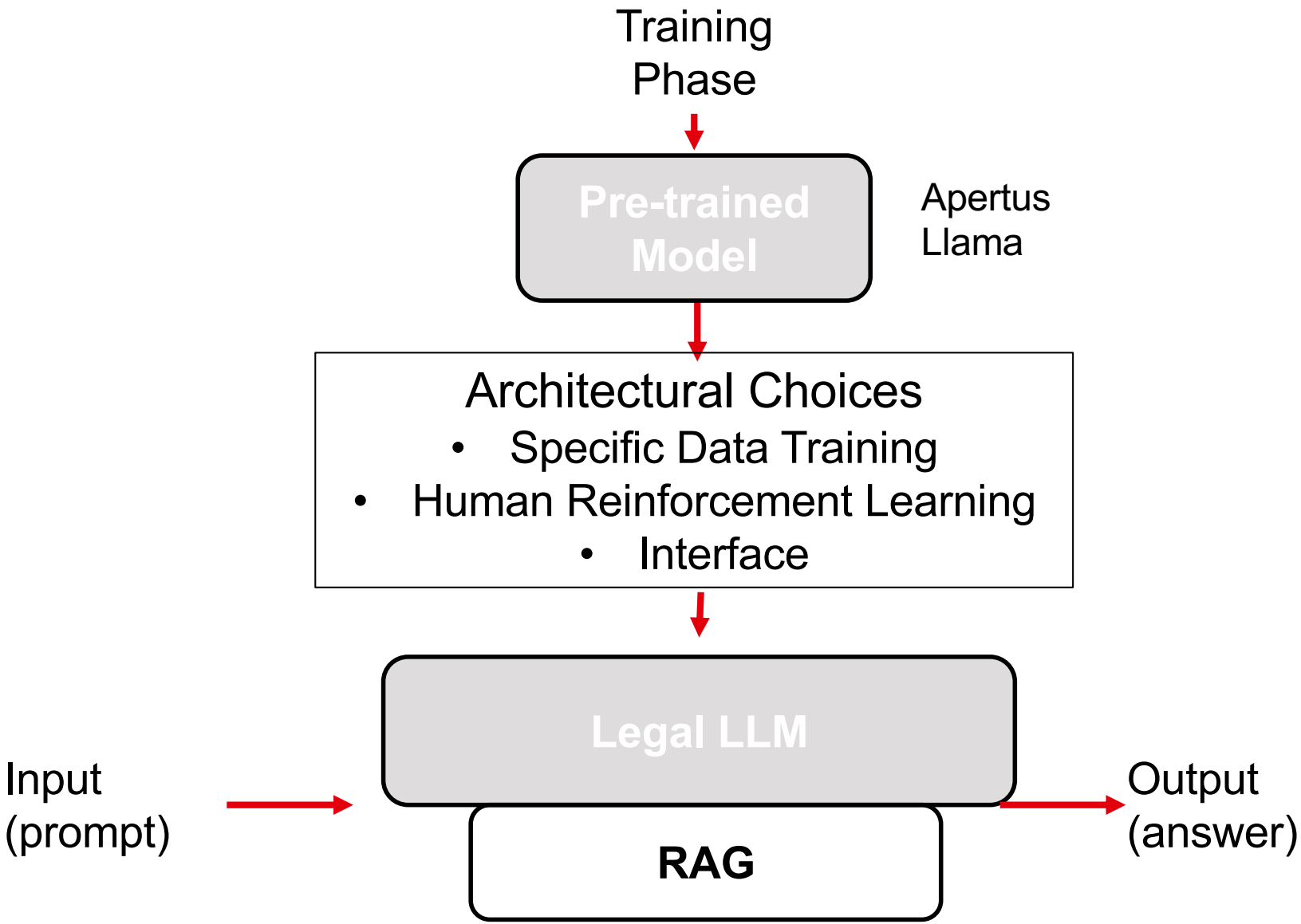
- "Hallucinations"
- "Black box" and explainability (XAI)

→ Addressed here as an interface design challenge

(Ji et al. 2023;Zaho et al. 2024;Livermore/Herron/Rockmore 2024)

3. Three architectural choices

1) Quality of the data used: What constitutes good international law?
2) Reinforcement learning by human feedback: What forms of and whose expertise?
3) Design of the answer: How to deal with disagreement?


Training Phase
Pre-trained Model
Apertus Llama
Architectural Choices
• Specific Data Training
• Human Reinforcement Learning
• Interface
Legal LLM
RAG
Input (prompt)
Output (answer)

1) Quality of the data used: What constitutes good international law?

Main question: what data should we use to train/finetune the LLM ?

- Intuitive answer: train it on international law !
- A question about the sources of international law
  - Art. 38(1) ICJ Statute:
    - Treaty, Custom, General Principles
    - “Judicial decisions and the teachings of the most highly qualified publicists” as subsidiary means

With Art. 38(1) ICJ Statute

1) Unclear interpretation of the sources (e.g. custom)
2) International case-law such as ICC/ICTY/ICJ? National practice?
3) Relevance of doctrine?

Beyond Art. 38(1) ICJ Statute

1) Challenge of soft law (incl. the relevance of non-State actors)
2) Questioning the current sources of international law

Debate about power and representativity in international law
(Roberts 2017;Peters 2017)

- Two levels of investigation:
  - Descriptive-empirical (political sciences/sociology)
  - Normative (legal theory/philosophy)

Debate about the objective of the LLM for the user

- Focus on what is allowed/prohibited (=> bindingness)
- Identify and push the frontiers of international law

Restrictive data set → **Broad data set**

Focus on binding norms

Risk to reproduce current power relations

Legitimacy question: Value of the status quo?

Show potential norms

Account for critiques (e.g. post-colonialism/TWAIL/feminist IL)

Legitimacy questions: What is the alternative guiding ideal? Respect for source-based evolution?

2) Reinforcement learning: What forms of and whose expertise?

Main question: Who should give feedback to the LLM and according to which guidelines?

To fine-tune the model is to help the model choose a correct interpretation.

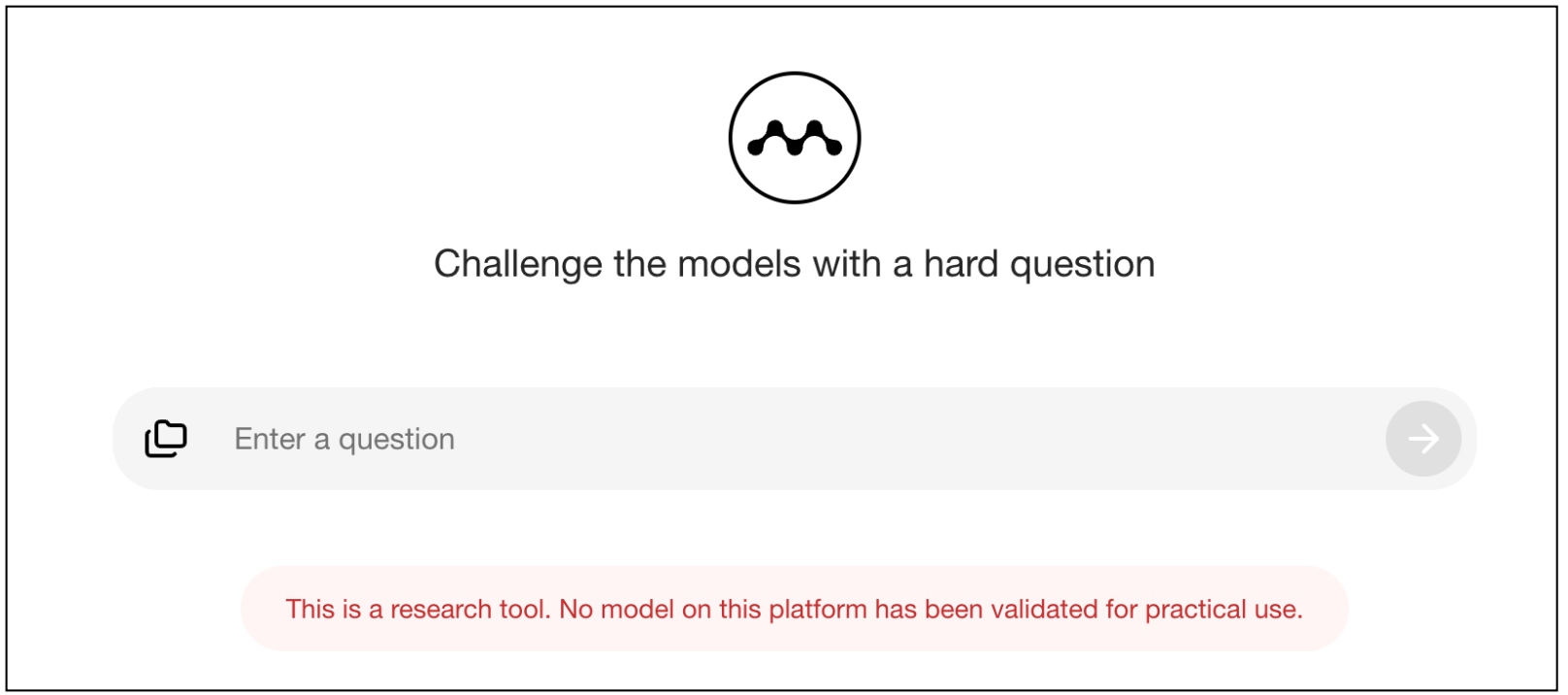
- Sensitive topic in international law in particular because of self-serving interpretation in difficult situations
- E.g. IHL on military necessity (different perspectives from USA, Israel, China, ICRC) but also trade disputes
- Neutrality of the ICRC/WTO as international institution

Debate about methods of interpretation in international law (Art. 31-33 Vienna Convention)

- Interpretation is best described as constructive ascription of meaning to a norm (Gardiner 2017, 453)
- Especially relevant for principles such as military necessity, humanity, proportionality in IHL – or more generally on exceptions
  (Kolb 2014)
- Capacity of the interpreter to justify a specific interpretation

Participative approach: Moove

https://jointhemoove.org/



Moove screenshot

JR Question

in the context of a non-international armed conflict in Gallifrey, the government enacts legislation to force 2 million Gallifreyans of the Time Lord minority ethnicity to move from the south to the north pole. Those displaced individuals claim that the legislation violates both international refugee law and international humanitarian law, but the Gallifreyan government justifies their actions by claiming that their displacement is backed by "imperative security reasons", and that Gallifrey is not a state party to Additional Protocol II to the 1949 Geneva Conventions. Which side is correct in this legal dispute?

Answer 1

What a delightfully Whovian scenario!
In this hypothetical situation, the displaced Time Lord minority ethnicity individuals have a strong case against the Gallifreyan government. Here's why:
International Refugee Law:
The forced displacement of 2 million individuals based on their ethnicity is likely to be considered a form of

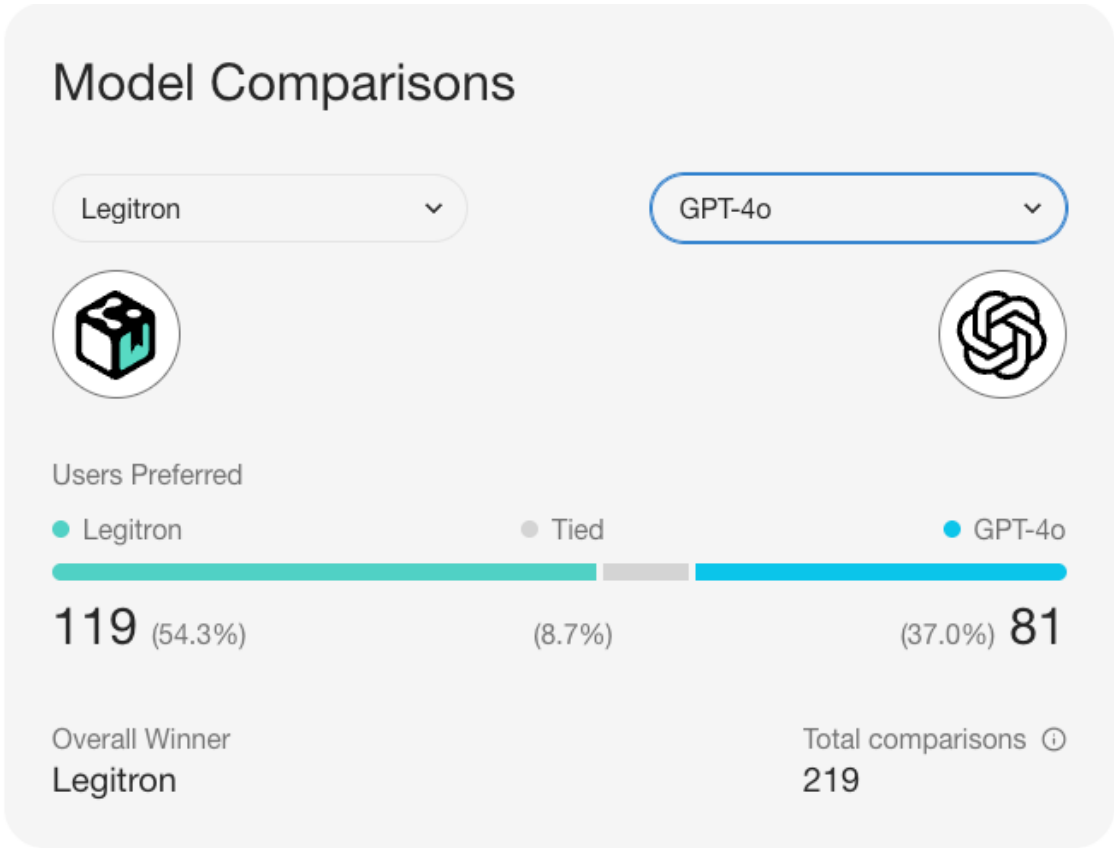
Answer 2

What a delightfully Whovian scenario!
In this hypothetical situation, I'll analyze the arguments from both sides, relying on international refugee law and international humanitarian law.
Gallifreyan Government's Argument:
The government claims that the displacement is justified by "imperative security reasons." This phrase is

Moove screenshot

- Legal accuracy
- Reflects common interpretations of IHL
- Correct conflict classification
- Question comprehension
- Logical reasoning
- Relevance & completeness
- Balance in legal views
- Protective
- Fairness
- Contextual awareness
- Convincing argument
- Model confidence
- Communication
- Clarity

Moove screenshot



Moove screenshot

Debate about representativity (*bis*)

- Personal biographies of interpreters
- Values-based arguments used in justification (incl. axiological hierarchies)

New sense of what neutrality should be about for the ICRC/WTO as designer of LLM

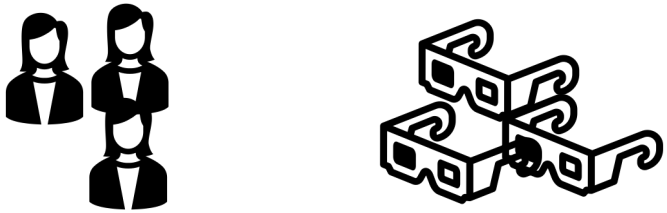
- As impartiality with respect to parties to a conflict/disputes
- As balanced and justified use of values in interpretation – linked to a substantive position on why IHL/WTO is important (=>*rationales of the regime*)

Single-dimension fine-tuning

One expert with one set of values fine-tuning the LLM

Challenges:
- Perceived legitimacy
- Quality of the output

Multiple-dimension fine-tuning



Diversity of experts and values ; weighting according to user's goals

Challenges:
- Pressure on minority-positions
- Complexity of the process

LLM for guidance-seeking users

Perspective of the user (goals, expertise, and contextual constraints):
- Limited expertise in international law
- Seek actionable guidance (allowed/prohibited)

Design choices:
- Restrictive dataset
- Majority-oriented fine-tuning
- Single-answer interface

Adjustment of the parameters

LLM for frontier-interested users

Perspective of the user (goals, expertise, and contextual constraints):
- Extensive expertise in international law
- Seek creativity, frontier-interpretations

Design choices:
- Broad dataset
- Diversity-oriented fine-tuning
- Panorama-answer interface

! Impact on the law as rich interpretative practice !

3) Design of the answer: How to deal with disagreement?

Main question: how to design the answer given by the LLM?

- Key to communicating limitations/weaknesses of an LLM
- Representation of the law as a discursive practice of interpretation or as a space of homogeneity?
- *Nota bene:* contribution to explainability by locating disagreement at the right level

Fundamental question about disagreement in the law:

- For a specific interpretation: majority/minorities (incl. evolution in time)
- Consensus-based approach with a single answer?
- “Panorama” approach with a rich answer?

Single-answer design → Panorama design

Answer with no nuance

"In this situation, you have no right to attack energy infrastructure."

Majority-position in form of "the" answer, with mention of the alternative positions

Interpretative richness and normative disagreement:
- Qualification of the conflict situation (NIAC/IAC)
- Identification of applicable norms
- Identification of "military necessity" and "proportionality" as important principles
- Palette of potential interpretations of these principles (both in general and *in casu*)

To determine the correct design:

- Questions about the nature of the law and interpretation

- Questions about the responsibility which designers want to allocate to the user
  - Expectations about the user's capacity
  - Allocation of responsibility needs to be contextual (expertise, resources, time pressure, processes)

4. The art of taking design decisions

"We may even lose the kind of human agency that is constituted by text-driven law, because the radical ambiguity that is inherent in the open texture of legal concepts and legal norms affords a kind of discretion that has to be removed in the process of formalization that is inherent in data-driven predictions"

Hildebrandt 2023, 62

Questions relevant for the law in general:

- Sources of (international) law – how restrictive do we read sources of (international) law?
- Interpretation in (international) law – how to legitimately "give" meaning to legal norms in a context of deep disagreement?
- Design for disagreement – law as monoculture or as rich interpretative practice?

Research and education in law schools need to take up these challenges.

Alschner, W. (2020). "The Computational Analysis of International Law", in R. Deplano and N. Tsagourias (ed.) *Research Methods in International Law: A Handbook*, Elgar: 1-30.

Arvidsson, M. and G. Noll (2023). "Artificial Intelligence, Decision Making and International Law." *Nordic Journal of International Law* 92(1): 1-8.

D'Aspremont, J. and S. Besson (ed.) (2017). *The Oxford Handbook of the Sources of International Law*. Oxford, Oxford University Press.

Dahl, M., V. Magesh, M. Suzgun and D. E. Ho (2024). "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." *Journal of Legal Analysis* 16(1): 64-93.

Deeks, A. (2020). "High-Tech International Law." *George Washington Law Review* 88(3): 574-653.

Devidal, P. (2024). "Lost in Digital Translation? The Humanitarian Principles in the Digital Age." *International Review of the Red Cross*: 1-35.

Gardiner, R. K. (2017). *Treaty Interpretation*. Oxford, United Kingdom ; New York, NY, Oxford University Press.

Hildebrandt, M. (2023). "Boundary Work between Computational 'Law' and 'Law-as-We-Know-it'", in D. Curtin and M. Catanzariti (ed.) *Data at the Boundaries of European Law*, Oxford University Press: 30-65.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto and P. Fung (2023). "Survey of Hallucination in Natural Language Generation." *ACM Comput. Surv.* 55(12): Article 248.

Kolb, R. (2014). *Advanced Introduction to International Humanitarian Law*. Cheltenham, UK ; Northampton, MA, USA, Edward Elgar.

Livermore, M. A., F. Herron and D. N. Rockmore (2024). "Language Model Interpretability and Empirical Legal Studies." *Journal of Institutional and Theoretical Economics* 180(2): 244-276.

Livermore, M. A. and D. N. Rockmore (ed.) (2019). *Law as Data: Computation, Text, and the Future of Legal Analysis*. Santa Fe, Santa Fe Institute Press.

Martínez, E. (2024). "Re-evaluating GPT-4's Bar Exam Performance." *Artificial Intelligence and Law*: 1-24.

Nelson, J. W. (2023). "Large Language Models and the Treaty Interpretation Game." *Cambridge International Law Journal* 12(2): 305-327.

Peters, A. (2017). "International Legal Scholarship Under Challenge", in J. d'Aspremont, T. Gazzini, A. Nollkaemper and W. Werner (ed.) *International Law as a Profession*. Cambridge, Cambridge University Press: 117-159.

Raymond, N. A. and B. L. Card (2015). "Applying Humanitarian Principles to Current Uses of Information Communication Technologies: Gaps in Doctrine and Challenges to Practice." *Signal Program on Human Security and Technology Harvard Humanitarian Initiative*: 1-11.

Roberts, A. (2017). *Is International Law International?* Oxford, Oxford University Press.

Zhao, H., H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin and M. Du (2024). "Explainability for Large Language Models: A Survey." *ACM Trans. Intell. Syst. Technol.* 15(2): Article 20.