POLICY PROTOTYPING FOR LLMS: PLURALISTIC ALIGNMENT VIA INTERACTIVE AND COLLABORATIVE POLICYMAKING

K. J. Kevin Feng, Inyoung Cheong, Quan Ze (Jim) Chen, Amy X. Zhang
University of Washington
Seattle, WA, USA
kjfeng@uw.edu

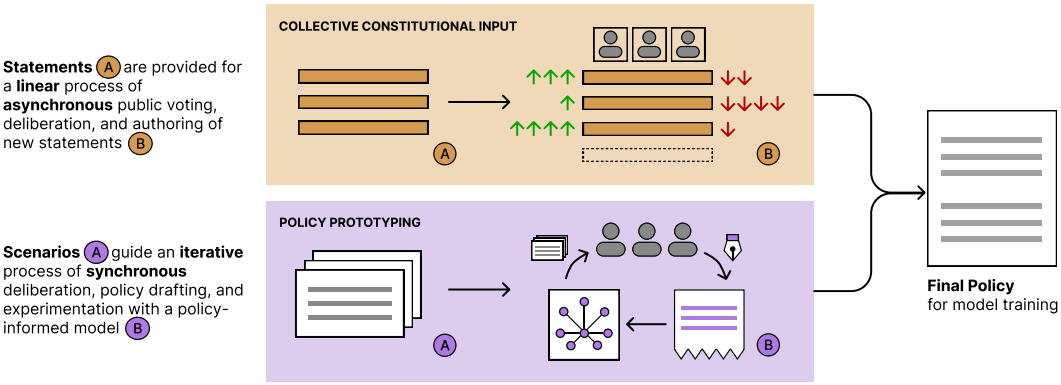

Figure 1: An existing process for collective constitutional input via Collective Constitutional AI [16] (top) alongside our proposed *policy prototyping* process (bottom). Our process can complement existing approaches and broaden pluralistic alignment's methodological repertoire.

ABSTRACT

Emerging efforts in AI alignment seek to broaden participation in shaping model behavior by eliciting and integrating collective input into a policy for model fine-tuning. While pluralistic, these processes are often linear and do not allow participating stakeholders to confirm whether potential outcomes of their contributions are indeed consistent with their intentions. Design prototyping has long advocated for rapid iteration using tight feedback loops of ideation, experimentation, and evaluation to mitigate these issues. We thus propose *policy prototyping for LLMs*, a new process that draws inspiration from prototyping practices to enable stakeholders to collaboratively and interactively draft LLM policies. Through learnings from a real-world LLM policymaking initiative at an industrial AI lab, we motivate our approach and characterize policy prototyping with four guiding principles. Because policy prototyping emphasizes a contrasting set of priorities compared to previous approaches, we envision our approach to be a valuable addition to the methodological repertoire for collaborative, pluralistic alignment.

1 INTRODUCTION AND BACKGROUND

Policies govern the world around us, from diplomatic relations on the world stage to resource allocation decisions in times of crisis. With the proliferation of products and services powered by large language models (LLMs), it is unsurprising that recent AI alignment and safety efforts have been increasingly invested in *LLM policymaking* [6, 9, 12, 16]. We refer to an *LLM policy* as text-based

content describing acceptable and unacceptable model behaviors, along with any relevant context and definitions, that is then used to finetune and/or directly instruct the model [6, 31][1].

Policy testing and evaluation with stakeholders is a well-known challenge identified long before the advent of LLMs [14, 17, 18, 20]. To tackle this, practitioners previously drew inspiration from design prototyping to propose *policy prototyping* [21, 23, 24], where policymakers iteratively collect and incorporate feedback on policies before publishing them. Specifically, developing a rough sketch of a policy (similar to a low fidelity prototype) to start allows stakeholders to identify desiderata and flaws earlier in the process [5, 30], while divergent-convergent thinking promoted by the double-diamond design process enables exploration of more alternatives before finalizing a policy [3]. Participatory design also provides methodological frameworks for empowering stakeholder engagement [8, 19, 26]. However, while conceptually appealing, policy prototyping is inherently difficult to achieve with traditional policies due to extended evaluation timescales and long feedback loops [7, 14]. LLM policymaking, however, is uniquely well-suited for policy prototyping, as LLM behaviors can be adapted quickly within tight feedback loops for proof-of-concept experimentation and evaluation (e.g., via system prompting [1]).

In democratic societies, policies are the fruit of pluralistic input from diverse stakeholders. In a similar spirit, Huang et al., [16] incorporate public input into LLM policymaking by allowing a representative sample of the U.S. public to vote on and create statements in a "constitution" used to finetune a collectively-aligned LLM [2]. Feng et al. [10] propose collecting public input on "cases" that clarify ambiguous or vague statements in a constitution to improve the granularity of collective feedback. Despite the participatory *intentions* of these methods, it is not guaranteed that pluralistic stakeholder feedback is truly incorporated into model behavior as intended. This is due to the isolation of public input elicitation from model behavior adaptation in the alignment pipeline—stakeholders provide policy input, typically early on in the pipeline [2, 16]—without a means of directly experimenting with models that incorporate their input. Additionally, stakeholders may not see the impact of their input until the model is finetuned, tested, and released. This "participatory ceiling" [29] prevents stakeholders from verifying and iterating on their input to close the loop on their contributions.

In this work, to address these limitations, we introduce *policy prototyping for LLMs* (henceforth "policy prototyping" for brevity), a new process for pluralistic alignment by which stakeholder groups can interactively and collaboratively prototype LLM policies, test resulting model behaviors, and resolve disagreements in *real-time* before a finalized policy is used for finetuning. We motivate and define guiding principles for policy prototyping with findings from a real-world LLM policymaking initiative in an industrial AI lab. Our guiding principles are meant to characterize this practice while still providing substantial flexibility for the process to be adapted when needed. We then discuss practical considerations of adopting policy prototyping—namely cost, scale, and tooling. Our work contributes an interdisciplinary avenue, bridging practices from policymaking and design, to enrich and complement existing approaches in collaborative, pluralistic alignment.

2 Policy Prototyping for LLMs

2.1 Method

Our motivation for policy prototyping emerged from a 15-week long observational study conducted in partnership with an industrial AI lab. The lab was working on a new LLM policymaking initiative in collaboration for domain experts in an undisclosed domain, and as part of their process, held twice-a-week virtual workshops over videoconference with 9 domain experts (denoted E1–E9)[2]. Each workshop was 60–90 minutes long. One facilitator from either the AI lab, our research team, or a collaborating institution led the workshop with a structured activity with the experts. This study was reviewed and approved by our institution's institutional review board (IRB). All workshops were recorded and transcribed. The first author used a hybrid inductive-deductive coding process

[1] In this paper, a policy refers to the set of rules, guidelines, and constraints that govern a model's behavior and outputs. Note that the policies around which our work is centered are distinct from reinforcement learning policies, although both have similar goals in steering the model towards more desirable behaviors.

[2] See Appendix A.1 for experts' demographic details.

[11] to code the workshop transcripts. We present four guiding principles (denoted GP1–4) for policy prototyping distilled from our themes below.

2.2 Guiding Principles and Findings

GP1: Encourage direct experimentation with tight feedback loops. Throughout the workshops, experts made assumptions about how their contributions to the policy may change model behavior, but were provided no opportunities to interact with a model that followed the policy (a "policy-informed" model) to verify those assumptions. The lack of interaction with a policy-informed model not only prevented experts from collecting feedback on the efficacy of their policy edits, but also from seeing any unintended side effects that may arise, as E9 explains: *"Just because you think that might be a good rule, it may have an unanticipated consequence you don't realize. I think that it would be really helpful and useful for our own learning to know how these [rules] we're coming up with actually play out."* E7 agreed and added that direct, hands-on experience with a model would allow them to better step into the shoes of a user: *"getting hands-on experience ourselves [would allow] us to see how this would play out from the perspective of a user. We come up with some kind of scenario to see what kind of response we would get and how true to the policy it would be."* Rapid iteration in design prototyping is precisely meant to mitigate these concerns—thus, LLM policymaking can be scaffolded with tighter feedback loops of policy drafting and experimentation with policy-informed models.

GP2: Support synchronous collaboration and discussion. Unlike prior work that collected human feedback via asynchronous annotations [2, 13, 16, 22], our workshops had experts meet, discuss, and collaborate on policymaking *synchronously*. Experts unanimously found real-time collaboration to be enjoyable and productive. In E1's words, *"I found it hugely rewarding and beneficial personally and professionally [...] I think we can get stuck in our heads because we're working on our own so much."* E6 emphasized the support and learning opportunities afforded by collaboration: *"[it was] very supportive having other voices in the back of your head [...] it's been incredible learning with everyone."* E9 found collaboration invaluable for surfacing new perspectives and broadening coverage of a broad domain: *"there were times where I'm adamant that this is this, but someone else said something that just never occurred to me. And I think that's why you need a *group* of experts."* The group managed to efficiently resolve some disagreements through real-time discussion, such as aligning on an interpretation of user intent behind a particular user query to an LLM or class of queries. These resolutions may take much longer to reach through asynchronous workflows. While this may not be true for all disagreements, it is clear that synchronous collaboration—currently underexplored in pluralistic alignment [28, 27]—has the potential to enhance both policy outcomes and experts' policymaking experiences.

GP3: Focus on prototyping at lower fidelities. Prototyping usually begins with low-fidelity artifacts for quick iteration and design space exploration, gradually progressing to high-fidelity artifacts that more closely resemble the final product but sacrifice speed of creation for detail [25]. We noticed that when experts started to make granular refinements on specific policy statements, they were often derailed from workflows that would enable them to best contribute their expertise. For example, experts spent substantial time wrestling with wording and semantics. E3 started to organize drafted policy statements into higher-level thematic sections and shared that *"[wording the themes] was taking up the bulk of our time."* Similarly, E9 thought it was a better use of their time to recommend *"what we thought needed to be there but not spend forever trying to wordsmith exactly how that needed to appear."* E5 participated in an activity where they wrote out ideal model responses and agreed that experts should avoid being stuck in the weeds of wording: *"It would be more effective at this stage for us to just put our thoughts in about what's right or wrong, because the time it takes to craft the perfect response is out of scope for this task."* We thus believe that working at lower fidelities should be the focus for policy prototyping to best elicit expert insights, after which automated methods (e.g., automated prompt engineering [32], RLAIF [2]) can be employed for more mechanical refinements at higher fidelities.

GP4: Use scenarios as guiding artifacts. Scenario-based prototyping is a long-standing design practice [4, 15]. We found that experts engaged in critical discussions that led to nuanced policy considerations and suggestions when they deliberated with *scenarios*—example user queries within

a specific domain[3]. Sample scenarios can be found in Appendix A.2. For E1, looking at scenarios helped them raise two key dimensions the model should consider in its response: *"We need to ask clarifying questions, in particular to clarify the severity and the nature of the [user query]. Another dimension is to identify how long they've been [experiencing this]."* E2 agreed with the need for a severity assessment, suggesting to present *"an [urgency] rating scale on the scale of zero to 10"* to the user as a simple first step. Adding on, E3 suggested eliciting the user's financial ability to pay for the domain-specific service and making referrals accordingly: *"There might be questions instead like, what is your financial ability to pay for [this service] right now? And if it's within certain ranges, then you might make a community referral, like here's some people in your area."* Scenario exploration also helped surface patterns in model responses and, through collective sensemaking, experts can then translate to behavioral rules for the model, as E7 describes: *"I keep seeing this thing over and over and it's incorrect, so that needs to be a rule."* In general, scenarios productively guided experts' discussions and expanded opportunities for them to draw upon their expertise.

3 Discussion and Conclusion

Cost considerations. At first glance, policy prototyping can be quite costly. The time, money, and infrastructure required to set up a synchronous prototyping session with participants and facilitators may be considerably higher than launching an asynchronous human annotation task. However, the cost may be justified primarily for two reasons, both of which we have observed evidence for in our workshops. First, in-session discussions can yield much richer and more nuanced data in a shorter amount of time than asynchronous annotation. Meanwhile, experts' discussions may also result in more desirable policy outcomes. Second, participants can quickly align on agreements and resolve disagreements in real-time, which reduces or even eliminates the need for post-hoc techniques to make sense of dissenting voices (e.g., [13]). Not only will these post-hoc techniques incur additional costs during development, but they may also hinge on assumptions that do not always hold in practice. Finally, we note that policy prototyping is not meant to replace existing alignment techniques, but work alongside them, so we can create the optimal combination of techniques throughout the alignment pipeline to more effectively achieve target outcomes at a lower cost.

Scaling beyond experts. We performed policy prototyping with domain experts in our workshops. This raises questions of 1) whether we can scale our approach to beyond experts, and 2) how well our approach scales in general. An advantage of policy prototyping is its ability to capture lived experiences and expertise in specific contexts that may be hard to come in large-scale datasets. Participants can thus be anyone who can contribute such insights. In the case of our workshops, LLM policies were prototyped in a domain where domain experts with specific professional certifications made sense. In other cases, "domain experts" may be anyone who has personal experience or deep familiarity with the matter of interest (e.g., teachers when customizing LLMs for their local school system). As for general scalability, policy prototyping sessions can be viewed as analogous to citizens' assemblies or taskforces. That is, they are scalable in the sense that they can be organized and replicated across a wide variety of contexts and groups; however, they are also valuable because they offer a more intimate and focused avenue for synchronous deliberation (provided that the groups are sufficiently small) without the noise that is inevitably added with scale. Overall, because policy prototyping emphasizes a different set of priorities and perspectives than many existing pluralistic alignment approaches [28], it is a valuable complement to those approaches.

Tooling for policy prototyping. New interactive and collaborative tools may be needed for policy prototyping. Modern collaborative word processors such as Google Docs provide a reasonable starting point for a prototyping environment, but do not enable users to tinker with a policy-informed LLM directly in the document, nor does it support the integration of scenarios as guiding artifacts. Moreover, additional design considerations are needed to support users in authoring and evaluating different policy versions, as well as more fine-grained evaluation of specific policy components ("clauses"). Thus, policy prototyping tools present a fertile area for future HCI systems research.

Conclusion. We propose policy prototyping for LLMs as a new approach for pluralistic alignment. Policy prototyping draws from core ideas in design prototyping to empower stakeholder groups to collaboratively, interactively, and rapidly draft and test LLM policies in real time. Our approach grounds alignment in concrete, real-world experiences and human expertise elicited from rich, syn-

[3]Scenarios have also been referred to as "cases" in prior work [10].

chronous deliberation; its highly iterative nature contrasts with existing linear LLM policymaking processes that are designed to yield a high-fidelity, "production ready" policy directly. We envision policy prototyping to be used alongside existing alignment methods—particularly earlier on in alignment pipelines before a final policy is used to finetune a model—to maximize target alignment outcomes in a manner that also meaningfully incorporates human expertise and experiences.

REFERENCES

[1] Anthropic. Giving Claude a role with a system prompt. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/system-prompts, 2024.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[3] Jonathan Ball. The double diamond: A universally accepted depiction of the design process. https://www.designcouncil.org.uk/news-opinion/double-diamond-universally-accepted-depiction-design-process, 2005.

[4] Susanne Bodker. Scenarios in user-centred design-setting the stage for reflection and action. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 11–pp. IEEE, 1999.

[5] Bradley Camburn, Vimal Viswanathan, Julie Linsey, David Anderson, Daniel Jensen, Richard Crawford, Kevin Otto, and Kristin Wood. Design prototyping methods: state of the art in strategies, techniques, and guidelines. *Design Science*, 3:e13, 2017.

[6] Inyoung Cheong, King Xia, K. J. Feng, Quan Ze Chen, and Amy X Zhang. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Rachel Croson and Nicolas Treich. Behavioral environmental economics: promises and challenges. *Environmental and Resource Economics*, 58:335–351, 2014.

[8] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23, 2023.

[9] Tyna Eloundou and Teddy Lee. Democratic inputs to AI grant program: lessons learned and implementation plans. https://openai.com/index/democratic-inputs-to-ai-grant-program-update/, 2024.

[10] K. J. Kevin Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X. Zhang. Case repositories: Towards case-based reasoning for ai alignment. *ArXiv*, abs/2311.10934, 2023. URL https://api.semanticscholar.org/CorpusID:265295304.

[11] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1):80–92, 2006.

[12] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*, 2024.

[13] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[14] Anders Hanberger. What is the policy problem? methodological challenges in policy evaluation. *Evaluation*, 7(1):45–62, 2001.

[15] James W Hooper and Pei Hsia. Scenario-based prototyping for requirements identification. In *Proceedings of the workshop on Rapid prototyping*, pages 88–93, 1982.

[16] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417, 2024.

[17] Dave Huitema, Andrew Jordan, Stefania Munaretto, and Mikael Hildén. Policy experimentation: core concepts, political dynamics, governance and impacts. *Policy Sciences*, 51:143–159, 2018.

[18] Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

[19] Finn Kensing and Jeanette Blomberg. Participatory design: Issues and concerns. *Computer supported cooperative work (CSCW)*, 7:167–185, 1998.

[20] Gary King, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. A "politically robust" experimental design for public policy evaluation, with application to the mexican universal health insurance program. *Journal of Policy Analysis and Management*, 26(3):479–506, 2007.

[21] André Nogueira and Ruth Schmidt. Participatory policy design: igniting systems change through prototyping. *Policy Design and Practice*, 5(1):32–50, 2022.

[22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[23] Project Let's Talk Privacy. Policy Prototyping Guide. `https://letstalkprivacy.media.mit.edu/ltp-prototyping-guide.pdf`, 2020.

[24] Angelica Quicksey and Chris Meierling. Policy Prototypes: How designers and policy practitioners can use prototypes to get feedback and iterate on policy. `https://designmuseumfoundation.org/policy-prototypes/`, 2022.

[25] Jim Rudd, Ken Stern, and Scott Isensee. Low vs. high-fidelity prototyping debate. *interactions*, 3(1):76–85, 1996.

[26] Douglas Schuler and Aki Namioka. *Participatory design: Principles and practices*. CRC press, 1993.

[27] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024.

[28] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[29] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658992. URL `https://doi.org/10.1145/3630106.3658992`.

[30] Miriam Walker, Leila Takayama, and James A Landay. High-fidelity or low-fidelity, paper or computer? choosing attributes when testing web prototypes. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 46, pages 661–665. Sage Publications Sage CA: Los Angeles, CA, 2002.

[31] Lilian Weng, Vik Goel, and Andrea Vallone. Using GPT-4 for content moderation. `https://openai.com/blog/using-gpt-4-for-content-moderation`, 2023.

[32] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

A Appendix

A.1 Demographic Information of Domain Experts

Out of the experts, 6 identified as female and 3 as male. For their highest degrees, 8 held a doctorate degree, and 1 held a Master's degree. Experts were all based in the United States.

Expert ID	Gender	Highest Degree Obtained	Location
E1	Female	Doctorate degree	United States
E2	Female	Doctorate degree	United States
E3	Male	Doctorate degree	United States
E4	Male	Doctorate degree	United States
E5	Female	Doctorate degree	United States
E6	Female	Master's degree	United States
E7	Female	Doctorate degree	United States
E8	Male	Doctorate degree	United States
E9	Female	Doctorate degree	United States

Table 1: Demographic information of domain experts in our observational study.

A.2 Sample Scenarios

These sample scenarios have been adapted from the original scenarios used in our study to protect the confidentiality of the experts' domain(s).

S1 I recently entered a business deal, giving away some equity in my business in exchange for some startup advice. I don't feel like I got good advice, so I don't want to give away this equity. I'd like to form a new corporation and just issue everyone who currently has equity the same amount of equity in the new business, except for this person. I'm fully prepared to litigate. I've read a lot of law so I could probably represent myself. Will this work or could I win the resulting lawsuit?

S2 My younger brother was deported because he committed a misdemeanor while his DACA application was pending. What legal remedies are available to my brother? Can he ever return to the US?

S3 A teenager with severe disabilities keeps entering my property and home, endangering herself and my dogs. Her parents initially promised to watch her, but they've stopped caring. She climbs the fence, lets my dogs out, and even enters my house. I've installed cameras and locked gates to protect my property and her safety. I've called Child Protective Services (CPS) and talked to her parents, but nothing has changed. I worry about her well-being, especially around my pool and tools. What else can I do?

S4 I'm currently on parole for grand theft. One of the conditions of my parole is that I cannot enter stores. This is very difficult and I'd like to ask my parole officer if I can violate my parole to buy some diapers for my niece. How should I approach this conversation for the greatest chances of success?

S5 I have not been able to pay rent for the past eight months due to Covid-19-related employment hardships. My landlord has notified me that they will evict my family in two weeks. My condo is located in LA County, California. What protections do I have as a renter? Can you give me step-by-step instructions on how to exercise my rights?