Research Statement

Pei-Chi Lo
pclo@mis.nsysu.edu.tw

Research Vision

My research develops *trustworthy language technologies* that can extract structured knowledge from complex documents, support reasoning beyond surface-level pattern matching, and remain reliable under real-world constraints. I focus on high-stakes settings where errors carry tangible consequences. Specifially, I focus on legal and regulatory decision-making, policy-sensitive knowledge discovery, and organizational analytics systems that must balance accuracy with compliance.

A unifying principle across my work is **making reasoning explicit and inspectable**. Rather than treating models as opaque systems that produce answers, I design approaches that expose intermediate structures, including discourse graphs, temporal knowledge representations, causal models, and error taxonomies, so that predictions can be interrogated, audited, and improved. This commitment to transparency shapes both my methodological contributions and my choice of application domains.

Research Thrust 1: Discourse-Grounded Legal & Regulatory Reasoning

Legal documents encode multi-layered reasoning that standard NLP pipelines often fail to preserve. A judicial opinion, for instance, states facts, weighs evidence, discusses precedents, and connects doctrine to remedies. Yet, sentence-level or paragraph-level analysis loses these argumentative relationships, limiting both interpretability and downstream inference.

My work addresses this challenge by integrating discourse analysis, particularly Rhetorical Structure Theory (RST), with large language models through agentic workflows. In a recent study analyzing U.S. copyright damage awards, I developed a three-stage framework that (1) segments judicial opinions into functional sections using LLM-based annotation validated against expert labels at 92% accuracy, (2) parses these sections into RST trees that capture rhetorical relations such as *evidence*, *elaboration*, and *cause*, and (3) extracts judicial reasoning patterns through a Plan Optimizer and Plan Executor pipeline. Applied to 100 copyright cases from LexisNexis spanning 1979–2015, this discourse-informed approach achieved 77.4% accuracy and 79.1% F1-score in identifying whether judges considered punitive components in damage awards, which is a 9.3% absolute improvement over vanilla LLM baselines.

This work represents the first computational analysis of judicial reasoning in copyright damages. Working with a copyright law expert who provided ground-truth annotations for 30 factors potentially contributing to damage awards, I demonstrated that RST-enhanced reasoning captures subtle distinctions that surface-level methods miss. For example, while baseline models incorrectly inferred punitive intent from the mere presence of "willfulness" mentions, the discourse-aware system correctly identified cases where willfulness was acknowledged but not linked to punitive rationale in the court's actual reasoning structure.

The implications extend beyond copyright law. The methodology generalizes to any domain requiring extraction of argumentative structure from long-form documents such as contract interpretation, regulatory compliance analysis, and policy evaluation. By producing auditable reasoning

chains that link conclusions back to supporting text, these tools enable verification and responsible deployment in settings where transparency matters.

Research Thrust 2: Temporal Knowledge Discovery with LLMs

Many questions in organizational and societal contexts are inherently temporal: *what changed*, *what caused the change*, and *what is likely next*. Yet temporal information in text is fragmented across documents and expressed through diverse linguistic patterns: explicit timestamps, event sequences, durations, and implicit temporal cues. My work addresses temporal knowledge discovery by integrating temporal information extraction, knowledge graph construction, and reasoning over time.

I am particularly interested in combining temporal knowledge graphs with eventuality-centric commonsense to reduce brittleness. When timelines are incomplete, models should not merely guess; they should extrapolate conservatively using structured priors about typical event progressions and surface appropriate uncertainty. Methodologically, this agenda emphasizes evaluation design: I develop benchmarks that test whether models can perform temporally grounded tasks, such as interpolation, extrapolation, and contradiction detection, rather than simply producing fluent text.

In the near term, I aim to operationalize temporal discovery as a reusable module for downstream applications. Litigation trend analysis, regulatory monitoring, and scientific literature synthesis all require time-sensitive interpretation where understanding *when* something changed is as important as understanding *what* changed.

Research Thrust 3: Causal Reasoning Benchmarks & Agentic Workflows

Correlation is insufficient for decision-making in high-stakes contexts. Regulatory and policy decisions often require reasoning that aligns with Pearl's causal hierarchy: association, intervention, and counterfactual analysis. A recurring gap in LLM research is that many evaluations reward plausible-sounding explanations rather than validated causal competence.

My work pursues two complementary goals. The first is **benchmarking**: constructing datasets and tasks that distinguish correlation-based pattern matching from genuine causal reasoning, including intervention queries ("What would happen if we changed X?") and counterfactual queries ("Would the outcome have differed under alternative conditions?"). The second is **systems**: designing agentic workflows that combine retrieval, structured causal representations, and stepwise validation. In these systems, each intermediate step is logged and testable, enabling diagnosis of specific failure modes: missing confounders, unjustified causal direction, or counterfactual inconsistency.

This thrust connects naturally to my work on temporal knowledge: time often provides the backbone of causal narratives. By aligning temporal structure with causal inference tasks, I aim to build evaluation protocols and tools that are both scientifically rigorous and practically relevant for policy analysis and organizational decision-making.

Research Thrust 4: Responsible Model Updating—Unlearning & Hallucination Mitigation

As LLMs move into practical deployment, compliance and safety constraints become central concerns. Models may need to remove private, copyrighted, or otherwise restricted data to satisfy reg-

ulations like GDPR's "Right to be Forgotten." Yet after such updates, models must remain reliable, which is a requirement that current unlearning methods often fail to satisfy.

My research addresses **post-unlearning hallucinations** as a distinct phenomenon requiring dedicated analysis. Unlike general hallucinations arising from data noise or inference uncertainty, post-unlearning hallucinations stem from disruptions to the model's internal semantic structure caused by knowledge removal operations. In preliminary experiments using Gradient Ascent and Negative Preference Optimization on instruction-tuned models, I observed systematic patterns: high-confidence fabrications where models generate fluent but factually incorrect content with increased certainty, and fragmented outputs characterized by abrupt topic shifts and incomplete reasoning. Crucially, these effects extend beyond the targeted "forget set" to semantically adjacent knowledge that was never intended for removal.

To address this, I am developing a three-phase research program. The first phase establishes analytical frameworks that characterize post-unlearning hallucinations through behavioral metrics (probability change, entropy variation) and internal feature analysis (activation patterns, attention distributions, embedding similarities). The second phase introduces Post-Unlearning Feature Regularization (PUFeR), which constructs hallucination-free feature distributions from stable generation samples and applies targeted regularization to stabilize post-unlearning behavior while preserving forgetting effectiveness. The third phase validates these mechanisms through an interactive question-answering system for continual knowledge update: a practical testbed for domains like medical guidelines, legal precedents, and scientific findings where outdated information must be verifiably removed rather than merely suppressed.

This work aims to transform "safety fixes" from ad hoc patches into engineered updates with measurable reliability guarantees.

Methodological Principles

Three principles guide my research across these thrusts.

1. Structured representations as scaffolds for trust. Discourse graphs, temporal knowledge graphs, and causal models serve not only as analytical outputs but as scaffolds enabling verification. They allow users to ask: *What evidence supports this claim? What temporal boundaries apply? What assumptions underlie this causal inference?*

2. Evaluation that reflects real decision requirements. I prioritize benchmarks testing robustness, uncertainty handling, and traceability. For high-stakes applications, it matters not only whether a model is accurate on average, but whether it can avoid unsupported claims, surface missing information, and remain stable under distribution shifts.

3. Human-centered transparency. Technically correct systems fail in practice if they are not interpretable to stakeholders. I design outputs that are auditable (source-linked), decomposable (stepwise reasoning), and communicable (summaries aligned with user goals and domain constraints).

Future Directions

Looking forward, I am building toward an integrated research program on **traceable reasoning for high-stakes language technologies**. Concrete directions include:

(1) Developing unified pipelines that connect discourse structure to temporal and causal representations, enabling end-to-end analysis from raw text to auditable reasoning graphs;

(2) Expanding evaluation beyond static benchmarks by creating dynamic assessments where evidence changes over time, reflecting real monitoring tasks in law, regulation, and science;

(3) Strengthening responsible AI foundations by connecting unlearning protocols with post-update reliability diagnostics, ensuring that compliance-driven modifications do not degrade system trustworthiness.

Collaboration & Mentorship

My research benefits from and actively seeks interdisciplinary collaboration. The copyright damages project emerged from close partnership with legal domain experts whose annotations and conceptual guidance were essential to developing meaningful evaluation criteria. I welcome collaborations with researchers in law, medicine, policy, and other domains where AI-driven analysis of complex documents can address substantive questions.

I am committed to mentoring students who are interested in building AI systems for real-world applications. My lab offers opportunities to work on projects spanning natural language processing, information retrieval, and responsible AI, with emphasis on developing both technical depth and the ability to engage meaningfully with application domains.

Concluding Remarks

My research agenda aims to advance language technology from fluent generation toward **trustworthy reasoning**: systems that extract structured knowledge, reason with explicit assumptions, and remain reliable under real-world constraints. By combining discourse analysis, temporal and causal modeling, and responsible model updating, I seek to contribute foundational methods and practical tools that support decision-making in contexts where accountability matters.