# Risk Radar: AI-Powered Detection of Unfair Contract Terms

Transfer Learning from European to Australian Law

Caitlin Douglas

# The Business Problem

# Consumer Risk & the Scalability Gap

## CONSUMERS

94% do not consistently Terms & Conditions [1]

Average T&C Length: 8,500 words, 20+ min to read [2]

Consumers suffer financial harm from unfair contract terms that contravene the Australian Consumer Law

## REGULATORS

ACCC AUSTRALIAN COMPETITION & CONSUMER COMMISSION

The ACCC has limited resources and cannot scale to protect all consumers

Consequently it is often reactive, not preventative

Manual legal review costs $150 - $200/ hour, making it impossible to audit every SME contract in the country.

# Risk Radar: a Demonstration

# Risk Radar: Gradio Interface

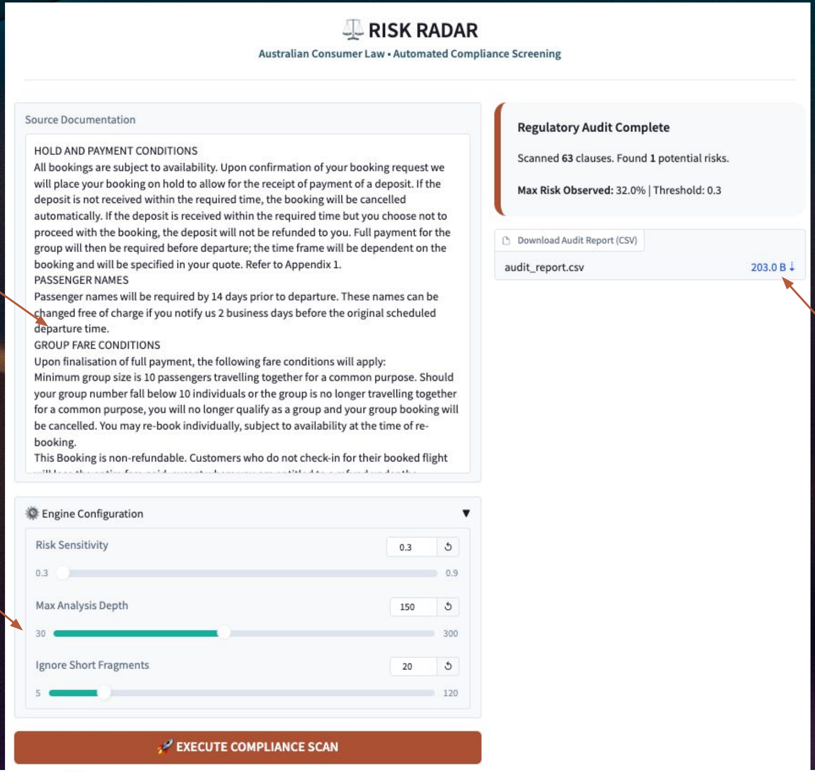
**Tier 4 Risk Radar Ensemble Model**

**Intended users:**

- Consumers
- Regulatory bodies
- Consumer advocacy groups
- Small businesses

Paste any or all clause from T&Cs

Adjust scan settings as desired




Clause-Level Risk Decomposition

| Triage | Outcome | Risk Score | Confidence | Method | Clause (full) |
|---|---|---|---|---|---|
| ⚠️ MED | Review Recommended (Ambiguous) | 0.32 | 0.68 | Tier 4 Ensemble | A change fee* applies for each passenger, for each flight segment changed. |

Instant legal analysis for unfairness

# Results that Drive Compliance

**64.7%**

**UNFAIR RECALL**

Reliability when flagging risky terms

**0.60**

**F1 SCORE (UNFAIR)**

Balancing our ability to catch unfair terms with the reliability of our detections

**93.8%**

**EFFICIENCY**

Reducing manual review from 20 minutes to < 1 second

# The Data

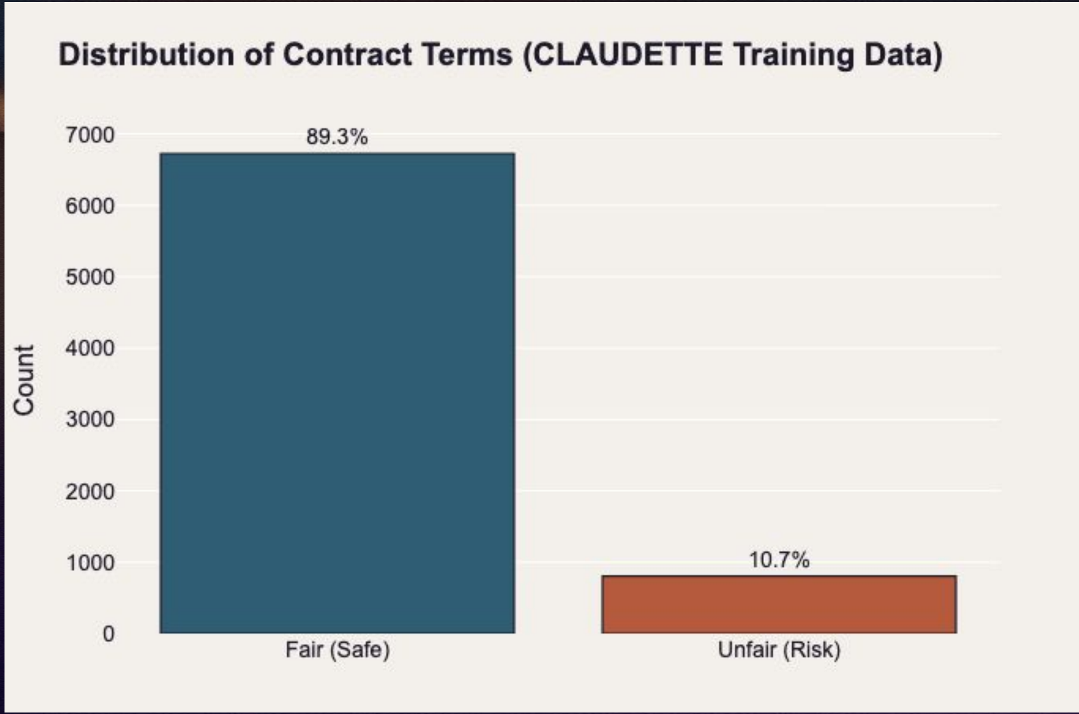
# The Data: European Claudette Dataset

**THE CHALLENGE**

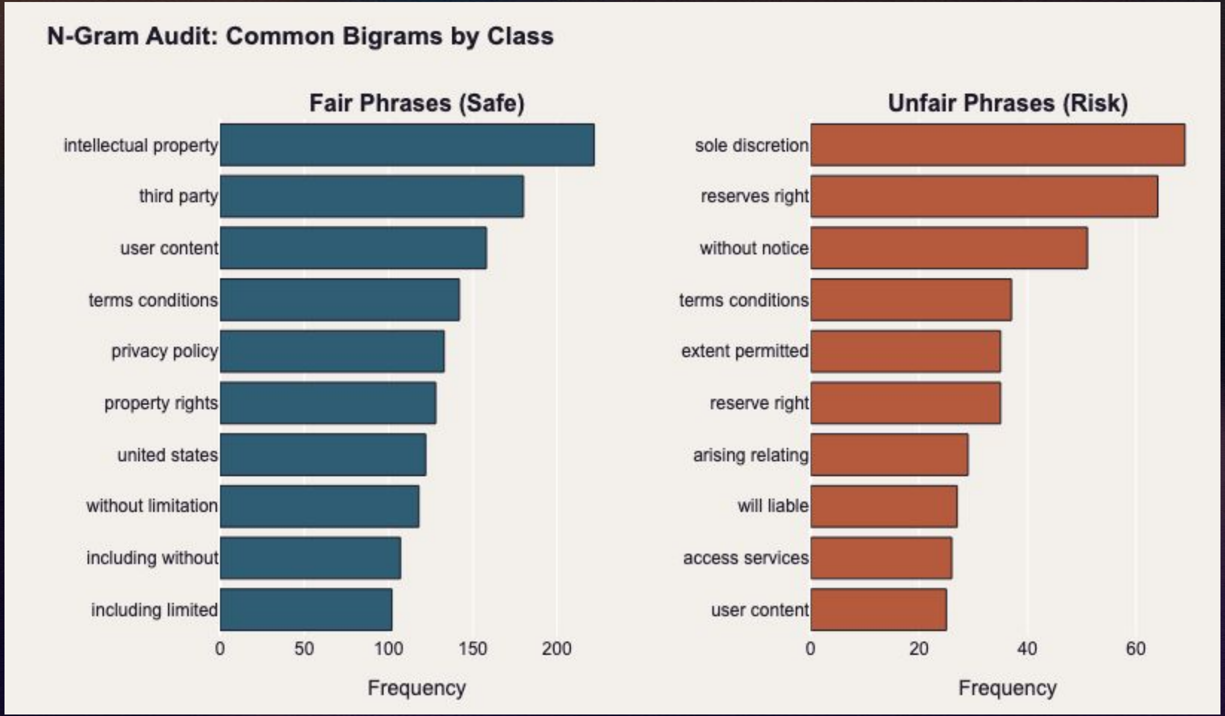
Australian Consumer Law (ACL) prohibits "unfair contract terms" (Sections 23-25), yet local training data is scarce

**THE SOLUTION: EU CLAUDETTE DATASET**

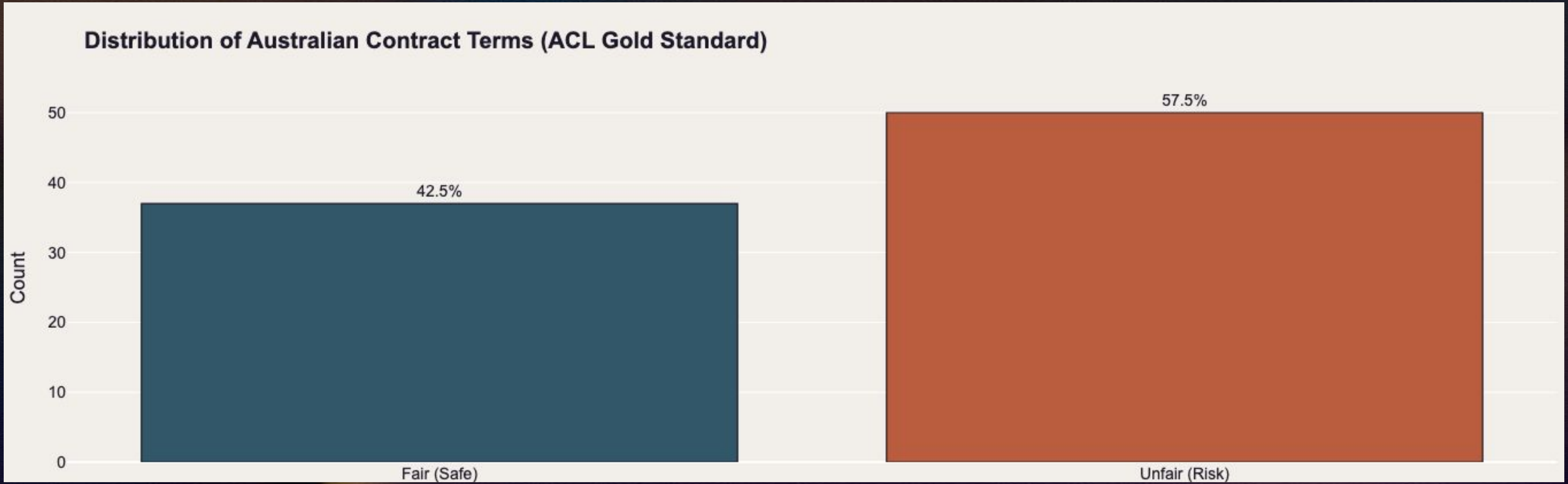
- 11,829 EU clauses split into fair/unfair classifications
- Severe Class Imbalance



# The Data: European Claudette Dataset



# The Test Data: Australian Dataset




**90+ fair and unfair terms and conditions derived from:**
→ Federal Court decisions on unfair contract terms
→ ACCC enforceable undertakings
→ ACCC regulatory guidance and published examples

# Methodology

**1. Benchmarking**
Compare performance of 4 models on Claudette Dataset

**2. Validation**
Review results against established literature

**3. Stress testing**
Evaluate the best-performing "European-trained" models on a raw Australian Dataset without retraining.

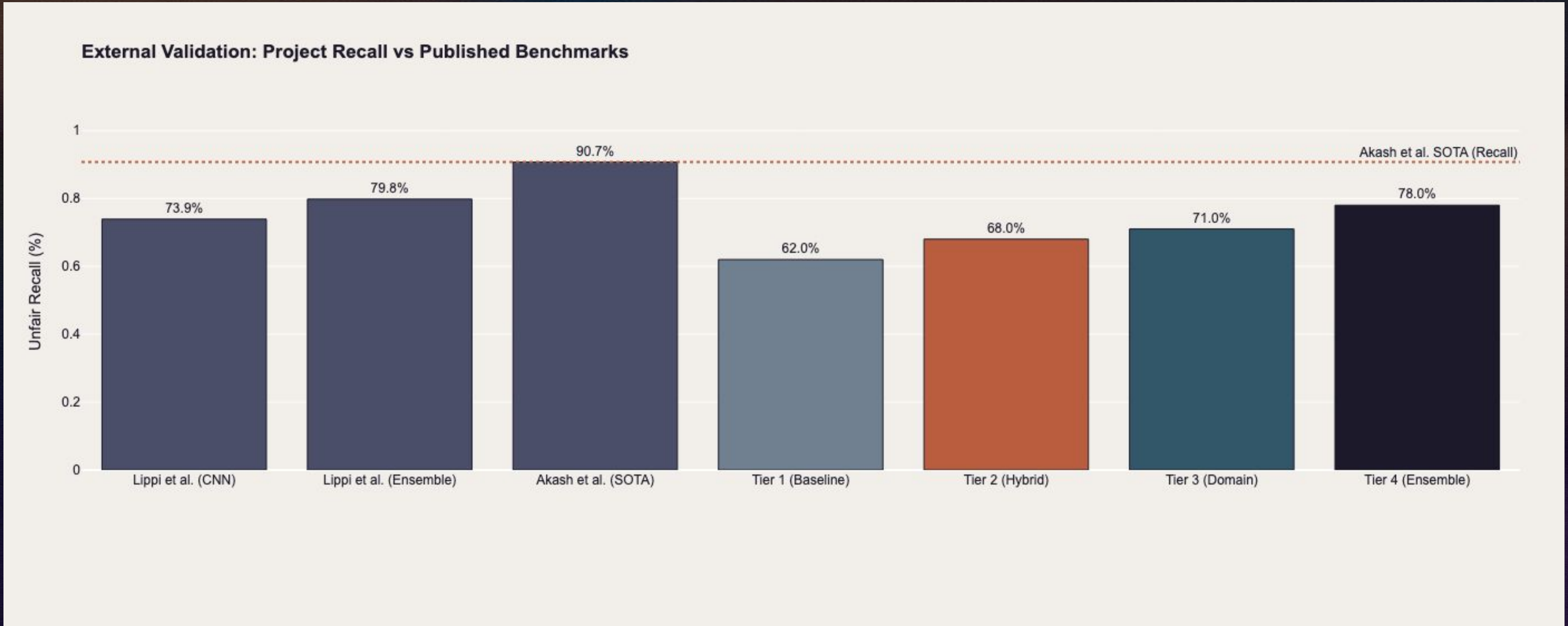
**4. Localisation**
Perform Few-Shot Fine-Tuning to recalibrate models for ACL nuances.

**5. Model selection for deployment**
Select best performing model for Gradio interface: Risk Radar.

# 1. Benchmarking the tiers on Claudette

| TIER | MODEL APPROACH | REGULATORY REASONING | RECALL SCORE (UNFAIR) | F1 SCORE (UNFAIR) |
|---|---|---|---|---|
| Tier 1: Baseline | TF-IDF + SVM | **Keyword spotter:** Looks for specific risky words, eg. *'sole discretion'*. | 62.0% | 0.73 |
| Tier 2: Hybrid | RoBERTa + SVM | **Semantic Screening:** Identifies 'grey-list' terms through contextual pattern recognition (flags too much). | 68.0% | 0.53 |
| Tier 3: Domain | Legal-BERT | **Domain Specialisation:** Resolves legal jargon and jurisdictional nuances using pre-trained legal logic. | 71.0% | 0.715 |
| **Tier 4: Ensemble** | **Risk Radar Ensemble** | **Risk Maximisation: Fuses lexical and semantic signals to ensure zero critical terms are missed.** | **78.0%** | **0.727** |

# 2. Benchmarking against Literature



# 3. The Real Test: Does it Work in Australia?

**The Challenge: Different legal jurisdictions**

- Trained on EU law
- Tested on ACL
- Only 90 test samples available

**ZERO SHOT (No Au training)**

**F1: 0.00**
**Recall 0%**

**Model couldn’t transfer directly**

→

**AFTER ADAPTATION (Fine-tuned on AU data)**

**F1: 0.68**
**Recall: 70%**

**Successfully adapted**

# Limitations and Next Steps

# Limitations

| | | |
|---|---|---|
| 01 | Small Australian Dataset | • Limited sample size constrains broad statistical confidence and high-resolution robustness estimates. |
| 02 | Contextual Isolation | • Clauses are analysed as independent units, ignoring broader document-level context and inter-clause dependencies. |
| 03 | Intentional False Positives | • A "Safety-First" bias leads to over-flagging, creating an administrative review burden. |

# Next Steps

**(1)**

**CONTEXTUAL DEPTH**

- Transition from isolated clauses to holistic contract evaluation.
- Expand classification to identify specific unfairness categories, eg indemnity, termination

**(2)**

**OPERATIONAL EFFICIENCY**

- Integrate direct regulator feedback to continuously refine model decision boundaries and reduce false positives.

**(3)**

**MARKET SCALING**

- Track industry-wide linguistic shifts in response to new Australian Consumer Law (ACL) reforms over time.
- Adapt the architecture to support non-English contracts, broadening coverage for diverse consumer bases.

# Thank you

# References

**Academic Peer-Reviewed Articles**

1. Akash, B. S., Kupireddy, A., & Murthy, L. B. (2024). Unfair TOS: An Automated Approach using Customized BERT. arXiv:2401.11207v2 [cs.CL].

2. Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law, 27(2), 117-139.

**Australian Regulatory & Industry Reports**

3. Consumer Policy Research Centre (CPRC). (2021). Submission to the Treasury: Enhancing protections against unfair contract terms. Treasury.gov.au. Retrieved from Treasury.gov.au.

4. Compare the Market. (2020). The Longest Terms and Conditions: Which websites take the longest to read? Retrieved from Compare the Market.