Gradient-Free Optimization of Retrieval Augmented Prompts for Legality Classification of Clauses in German Employment Contracts

David Pauschert

4th of March, 2025, Master's Thesis – Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Background & Motivation

Problem Statement & Research Questions

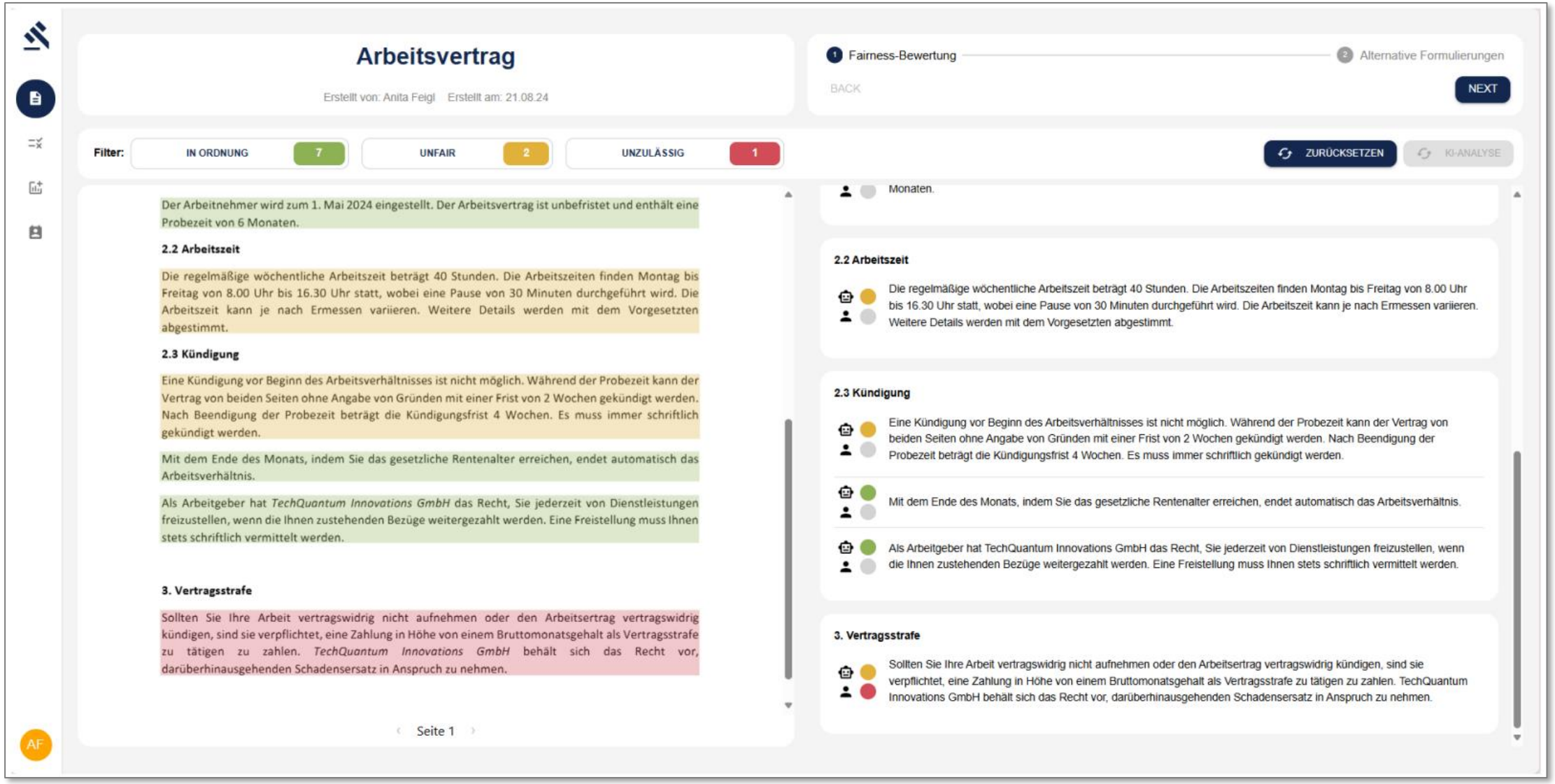
Methodology

Results

Discussion & Outlook

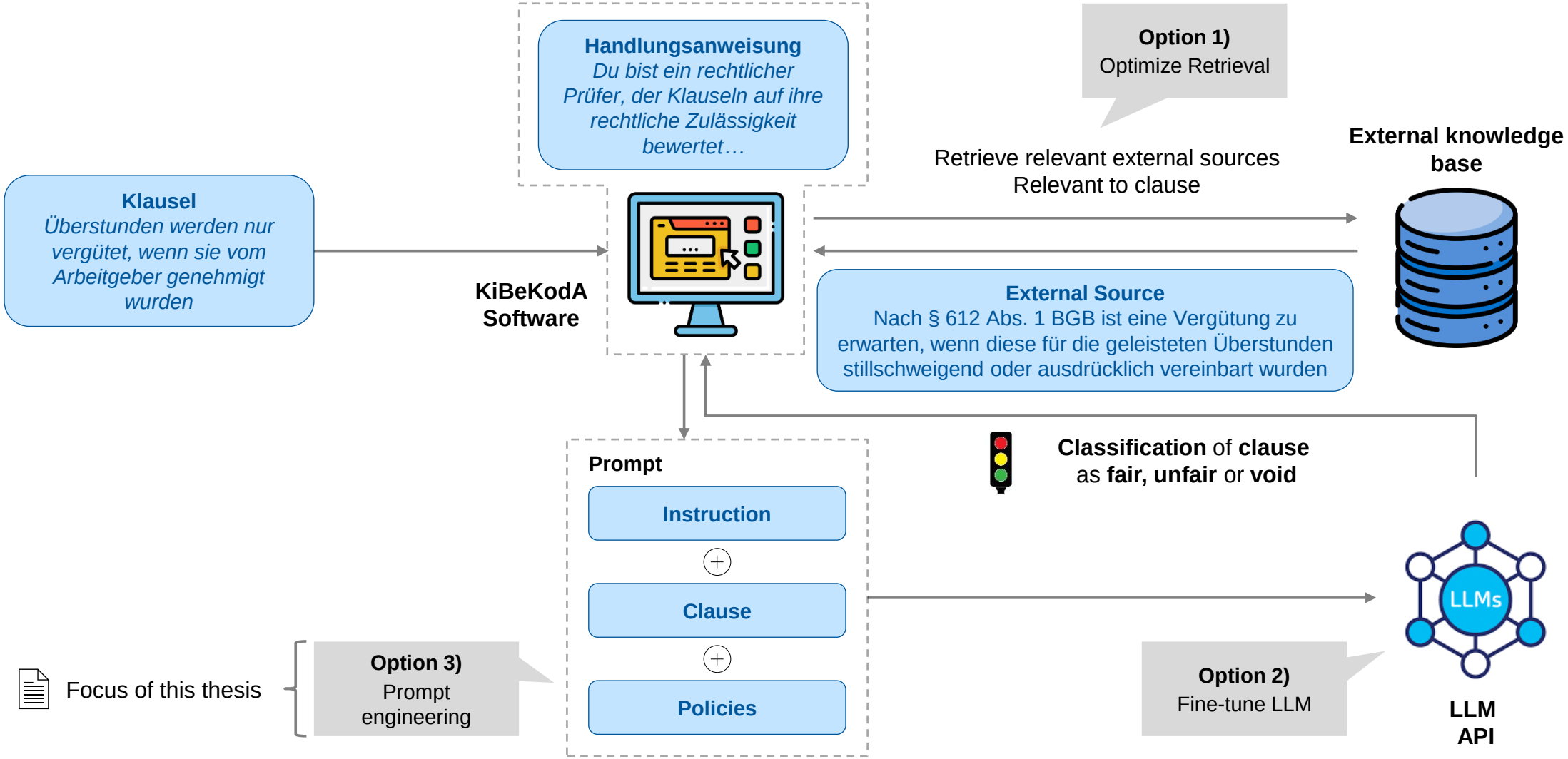
© sebis

KIBeKodA's aim is the development & evaluation of a that assists in the legal assessment and correction of German employment contract clauses



© sebis

The master's thesis aims to optimize the AI-Assisted Legal Analysis and Correction of German Employment Contracts



© sebis

Outline

Background & Motivation

Problem Statement & Research Questions

Methodology

Results

Discussion & Outlook

© sebis

The overarching research question is whether APO can help to learn the legal domain knowledge needed to accurately classify clauses

	1 RQ 1	2 RQ 2	3 RQ 3
Research question	What is the **current state** of **Automatic Prompt Optimization**, and to what extent do existing approaches consider **expert knowledge** in prompt optimization?	How can **existing methods** be **adapted** to design an **APO algorithm** that **optimizes prompts** used for **classifying clauses** from German contracts by implicity **learning** the **required legal expert knowledge**?	How does this **new APO algorithm perform** in classifying the fairness of clauses from German employment contracts **compared** to an **unoptimized expert-written prompt** as a **benchmark** using the Wardas-Matthes [x] dataset?
Method	Literature Review	Algorithm development	Benchmarking

© sebis

Outline

Background & Motivation

Problem Statement & Research Questions

Methodology

Results

Discussion & Outlook

© sebis

Dataset, Task & APO Objective

Background

893
Clauses

Extracted from contracts from law-firms client data

Each clause comes with label, category, title and explanation[1]

Label Explanation

Fair

Neither unfair nor void (78%)

Unfair

Major disadvantage for one party but not void (9%)

Void

Fails to comply with regulation or void under existing law (13%)

Clause Categories (Classes)

	# Clauses	# Policies	Fair / Unfair / Void
Termination	92	0	87 / 9 / 4
Penalty Clause	117	5	70 / 12 / 18
Statute of Limitations	30	0	53 / 7 / 40
Compensation	127	4	67 / 6 / 27
Format	24	1	46 / 29 / 25
Cut-off Periods	9	5	0 / 11 / 89
Illness	70	1	81 / 7 / 12
Garnishment	20	2	40 / 5 / 55
Execution	108	0	89 / 10 / 0
Tasks	87	0	91 / 3 / 6
Vacation	59	0	78 / 17 / 5
Overtime	8	2	75 / 12.5 / 12.5
Other	142	2	92 / 7 / 1

The dataset comes with **20 policies** written by lawyers to assist in **identifying void clauses**

1) Only if clause is unfair or void

© sebis

The current baseline prompt used to classify clauses is static, category-agnostic and consists of an instruction, policies and the clause

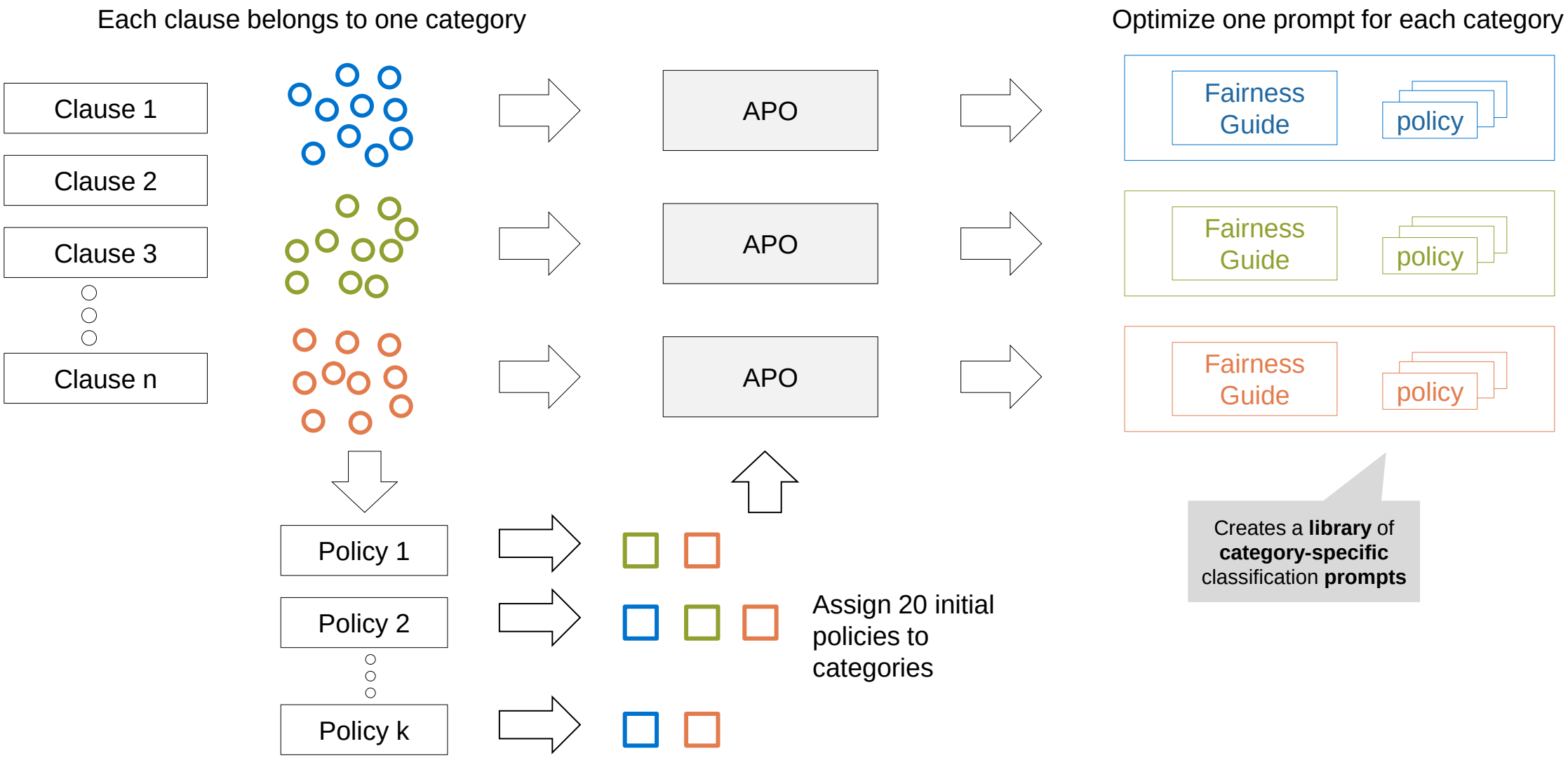
Baseline classification prompt template

```
Du bist ein deutscher Anwalt, welcher auf Wirtschaftsrecht spezialisiert ist und Klauseln aus
Arbeitsverträgen hinsichtlich ihrer rechtlichen Zulässigkeit und Fairness bewertet.

Du denkst Schritt für Schritt und beachtest dabei die Regeln, welche dir zusammen mit der
Klausel zur Verfügung gestellt werden.

Die Regeln stammen von professionellen Anwälten und sind somit immer als korrekt anzusehen.

Du bewertest eine Klausel nur dann als 'unzulässig', wenn sie gegen eine der gegebenen Regeln
verstößt.

Wenn keine der gegebenen Regeln Anwendung findet, nimmst du eine eigene Einschätzung vor.
Dabei bewertest du die Klausel aber nur entweder als 'unfair' oder 'unbedenklich'. Du
bewertest die Klausel als 'unfair' wenn sie den Arbeitgeber oder den Arbeitnehmer unangemessen
benachteiligt oder besonders ungewöhnliche Formulierungen enthält.

Deine Antwort beendest du immer mit einer abschließenden Bewertung als: 'unzulässig', 'unfair'
oder 'unbedenklich’.

Ist die folgende Klausel unzulässig, unfair oder unbedenklich?

Klausel: {clause}

Beachte bitte folgende Regeln: {rules}
```

Same static prompt used for classifying all 893 clauses

© sebis

Our APO algorithm is designed to find an optimal prompt for each category within the clause corpus



© sebis

For tuning prompts on category-level, we needed to restructure the prompt template – category prompt is created dynamically at runtime

Structure elements

Instruction

Policies

Fairness Guide

Response format

Clause as input

```
<anweisung>
Du bist ein deutscher Anwalt, spezialisiert auf Wirtschaftsrecht, und bewertest Klauseln der Kategorie {category} aus Arbeitsverträgen
hinsichtlich ihrer rechtlichen Zulässigkeit und Fairness.

Arbeite strikt Schritt für Schritt und beachte die Regeln, die dir zusammen mit der Klausel zur Verfügung gestellt werden. Diese Regeln wurden
von professionellen Anwälten erstellt und sind immer als korrekt anzusehen.

<vorgehen>
1. Prüfe im ersten Schritt, ob der Tatbestand von einer oder mehrerer der Regeln zutrifft. Falls ja, wende die Regeln an und wäge sie ab, um die
entsprechende Klassifikation (‚unbedenklich‘, ‚potentiell unzulässig‘ oder ‚unzulässig‘) der Klausel zu bestimmen.

2. Falls keine der Regeln zutrifft oder keine Regeln existieren, wende die bereitgestellten "Allgemeinen Prinzipien" zur Klassifikation an. In
diesem Fall sollst du die Klausel nur entweder als 'unbedenklich' oder 'potentiell unzulässig' nicht aber als 'unzulässig' klassifizieren.
</vorgehen>

<regeln>
{policies}
</regeln>

<Allgemeine Prinzipien>
{fairness guide}
</Allgemeine Prinzipien>

Befolge bei deiner Antwort bitte das gegebene Antwortformat und gib an, welche Regeln du angewendet hast, um die Klausel zu klassifizieren.
</anweisung>

<antwortformat>
      <bewertung>Gib hier die abschließende Klassifikation ein: {classifications}.</bewertung>
      <regeln>Gib hier durch Kommas getrennt ausschließtlich die IDs derjenigen Regeln an, die du angewendet hast, um die        Klausel zu
klassifizieren. Wenn keine der Regeln Anwendung findet, also die allgemeinen Prinzipien angewendet   wurden, dann bleibt die Liste leer
      </regeln>
      <erklaerung>Gib hier eine kurze Erklärung an, wie du zu deiner Bewertung gekommen bist</erklaerung>
</antwortformat>

<frage>Ist die folgende Klausel ‚unbedenklich‘, ‚potentiell unzulässig‘ oder ‚unzulässig‘?<frage>

<klausel>{clause}</klausel>
```


© sebis

We use an LLM-based optimization setup in which the LLM is used for all stages of prompt optimization process

Setup of simplified Demo

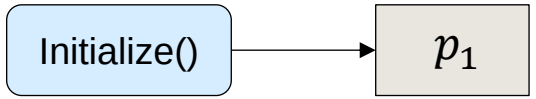
30 Clauses

Category **Termination**

9 void clauses, **3 unfair** clauses, **18 fair** clauses

No initial polices provided by lawyers

Simplified APO Process



© sebis

We initialize the prompt optimization process with a candidate consisting of a domain-expert written component and no policies

Initial candidate

Initial Candidate

Fairness Guide

Du bewertest die Klausel als 'potentiell unzulässig' wenn sie den Arbeitgeber oder den Arbeitnehmer unangemessen benachteiligt oder besonders ungewöhnliche Formulierungen enthält.

+

Policies

No Inital Policies for Category Termination

© sebis

We use an LLM-based optimization setup in which the LLM is used for all stages of prompt optimization process

Setup of simplified Demo

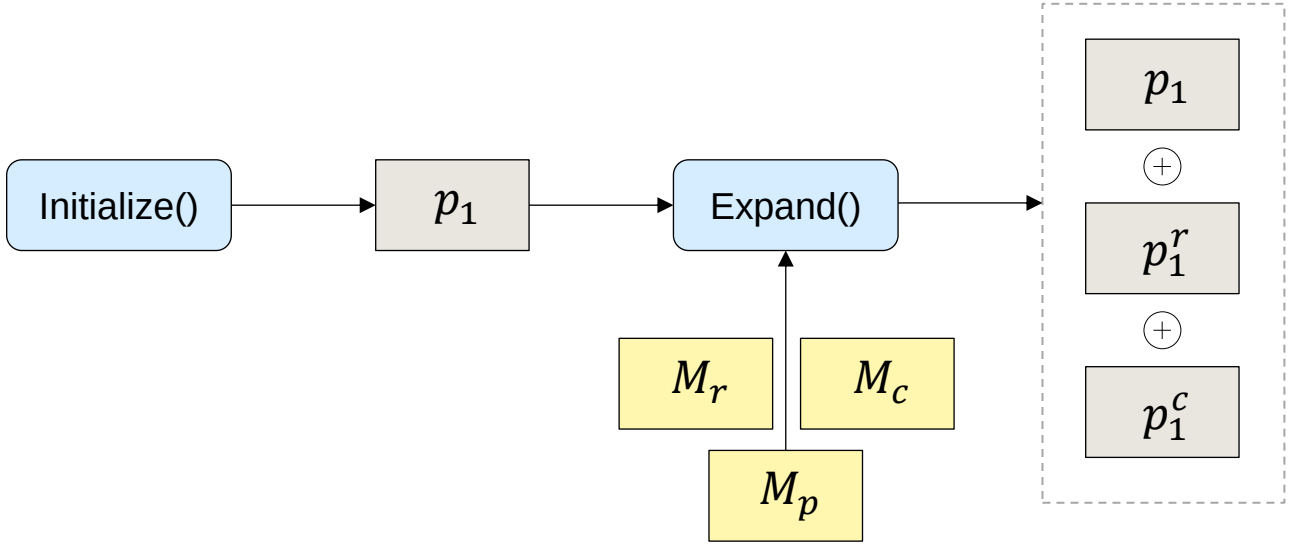
30 Clauses

Category **Termination**

9 void clauses, **3 unfair** clauses, **18 fair** clauses

No initial polices provided by lawyers

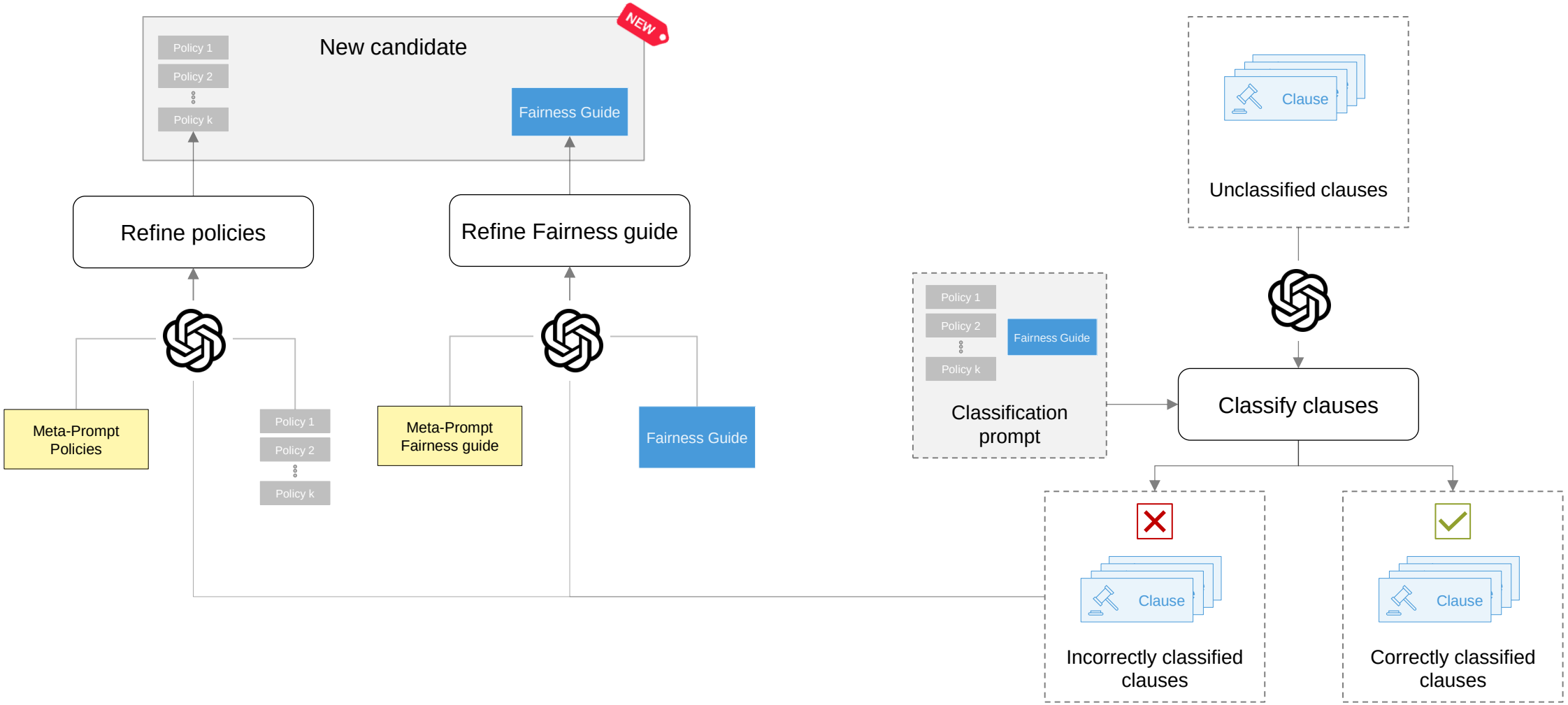
Simplified APO Process



© sebis

In each expansion step, both the fairness guide and the policies undergo a refinement process to create a new candidate

Prompt refinement process



© sebis

We initialize the prompt optimization process with a candidate consisting of a domain-expert written component and no policies

Refined candidate

Refined Candidate

Fairness Guide

Eine Klausel ist als 'unbedenklich' zu betrachten, wenn sie sowohl den Arbeitnehmer als auch den Arbeitgeber nicht unangemessen benachteiligt und im Einklang mit den gesetzlichen Standards sowie allgemeinen Prinzipien steht.
...

Eine Klausel ist 'potentiell unzulässig', wenn sie aufgrund ihrer Formulierung die Interessen einer Partei signifikant schlechterstellt, oder wenn durch die Regelungen Mehrdeutigkeiten oder potenzielle Benachteiligungen entstehen könnten, ohne klare, rechtlich fundierte Rechtfertigung.
...

+

Policies

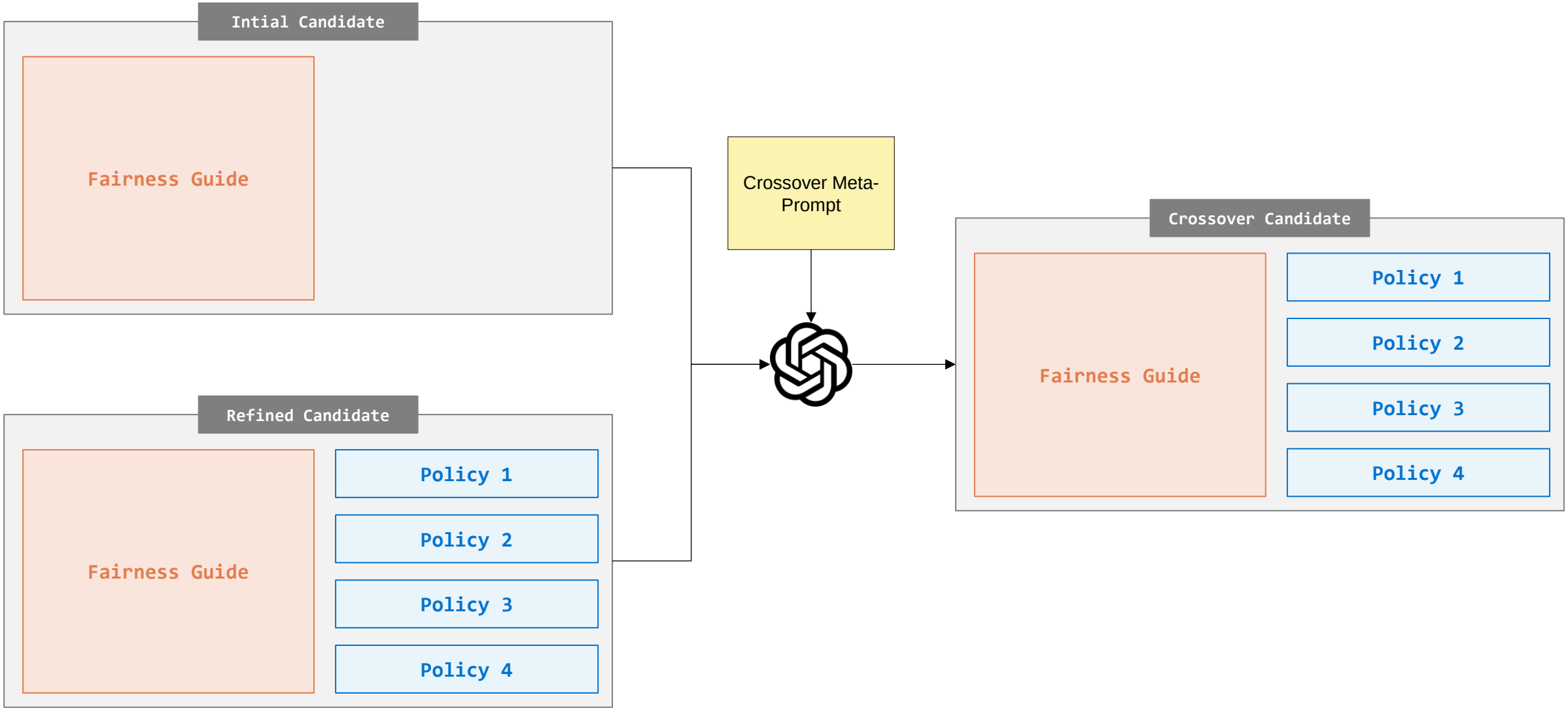
Eine Klausel in einem Arbeitsvertrag, die während der Probezeit nur eine Kündigung zum Ende eines Kalendermonats erlaubt, ist unzulässig. Es muss möglich sein, täglich mit einer Frist von 14 Tagen zu kündigen.

4 derived policies in total

Eine Klausel, die die Möglichkeit der täglichen Kündigung während der Probezeit auf bestimmte Tage im Monat einschränkt, ist unzulässig.

© sebis

Global exploration – we cross over two prompts in each update step to increase diversity in the prompt search space



© sebis

We use an LLM-based optimization setup in which the LLM is used for all stages of prompt optimization process

Setup of simplified Demo

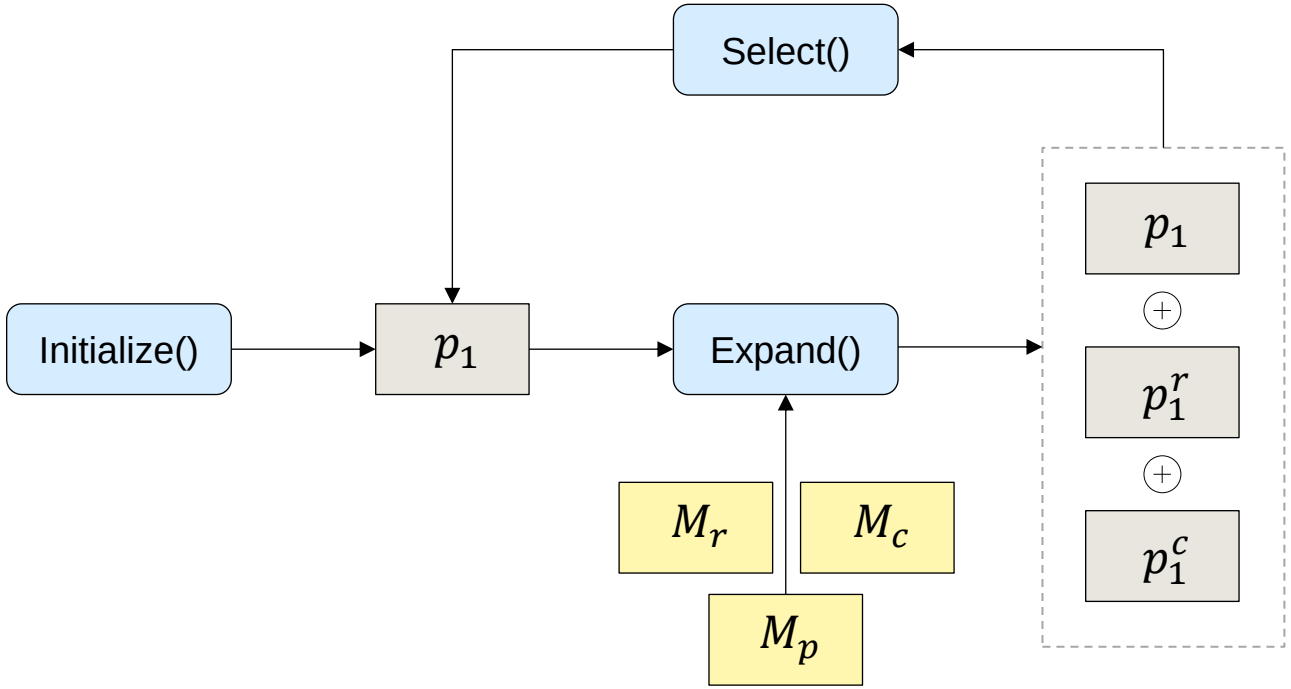
30 Clauses

Category **Termination**

9 void clauses, **3 unfair** clauses, **18 fair** clauses

No initial polices provided by lawyers

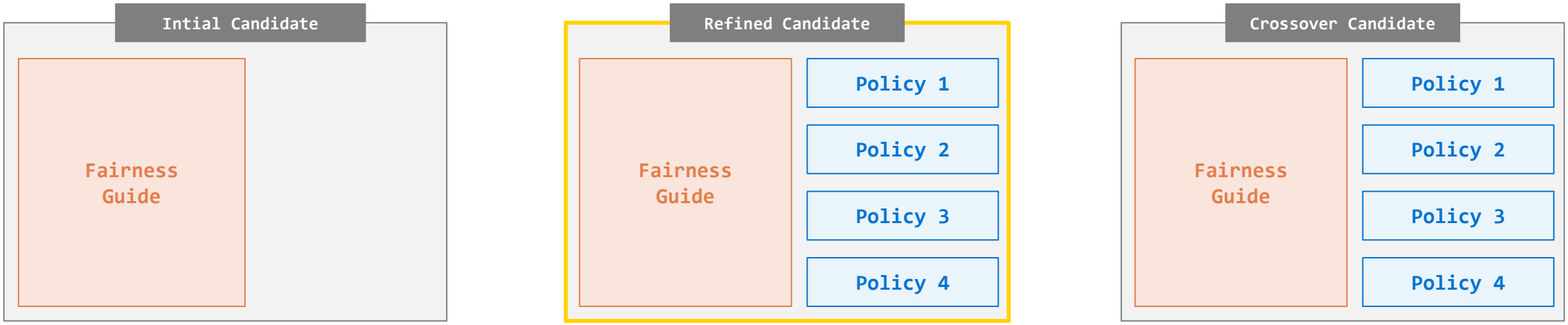
Simplified APO Process



© sebis

In each iteration, we select the best performing candidates to proceed to the subsequent APO iteration

Prompt selection process



F1: 0.36

		P		
		Fair	Unfair	Void
T	Fair	8	10	0
	Unfair	1	7	1
	Void	1	3	0

F1: 0.50

		P		
		Fair	Unfair	Void
T	Fair	9	4	5
	Unfair	1	3	4
	Void	0	0	4

F1: 0.46

		P		
		Fair	Unfair	Void
T	Fair	9	4	5
	Unfair	1	2	5
	Void	0	0	4

© sebis

We use an LLM-based optimization setup in which the LLM is used for all stages of prompt optimization process

Setup of simplified Demo

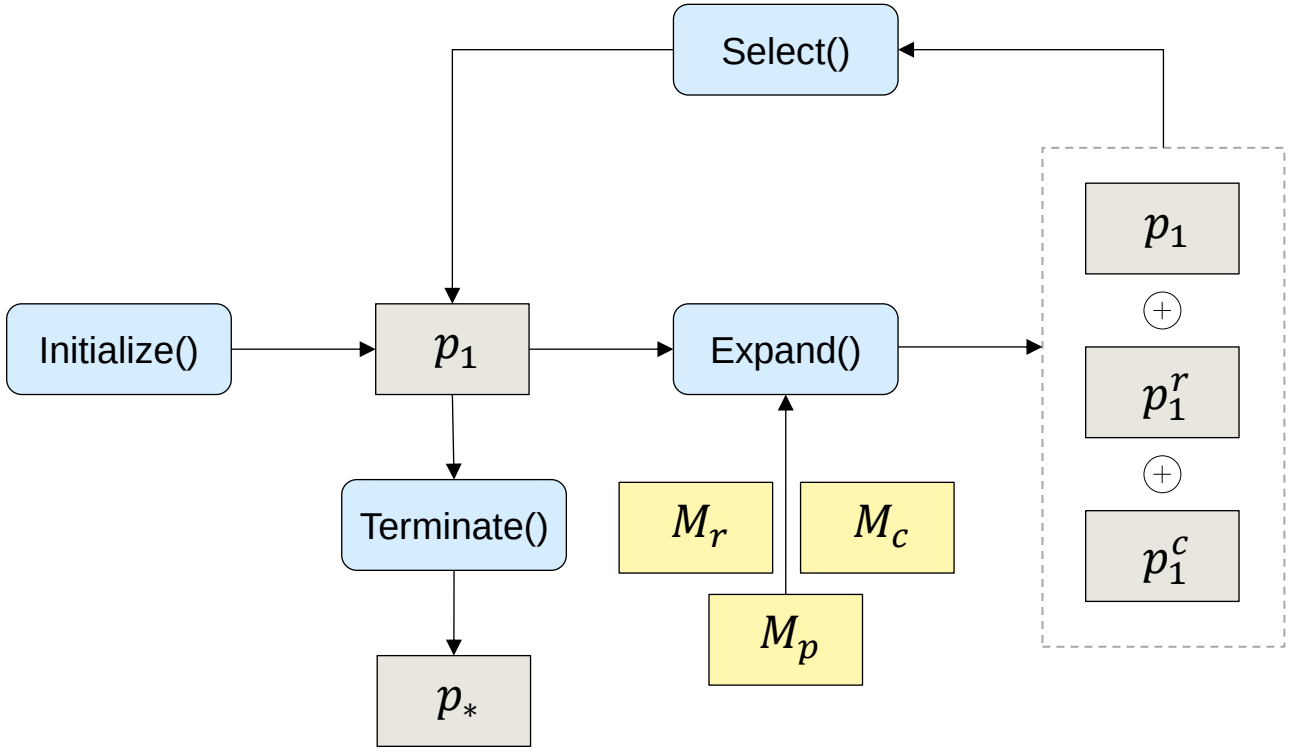
30 Clauses

Category **Termination**

9 void clauses, **3 unfair** clauses, **18 fair** clauses

No initial polices / examination guides provided by lawyers

Simplified APO Process



© sebis

Outline

Background & Motivation

Problem Statement & Research Questions

Methodology

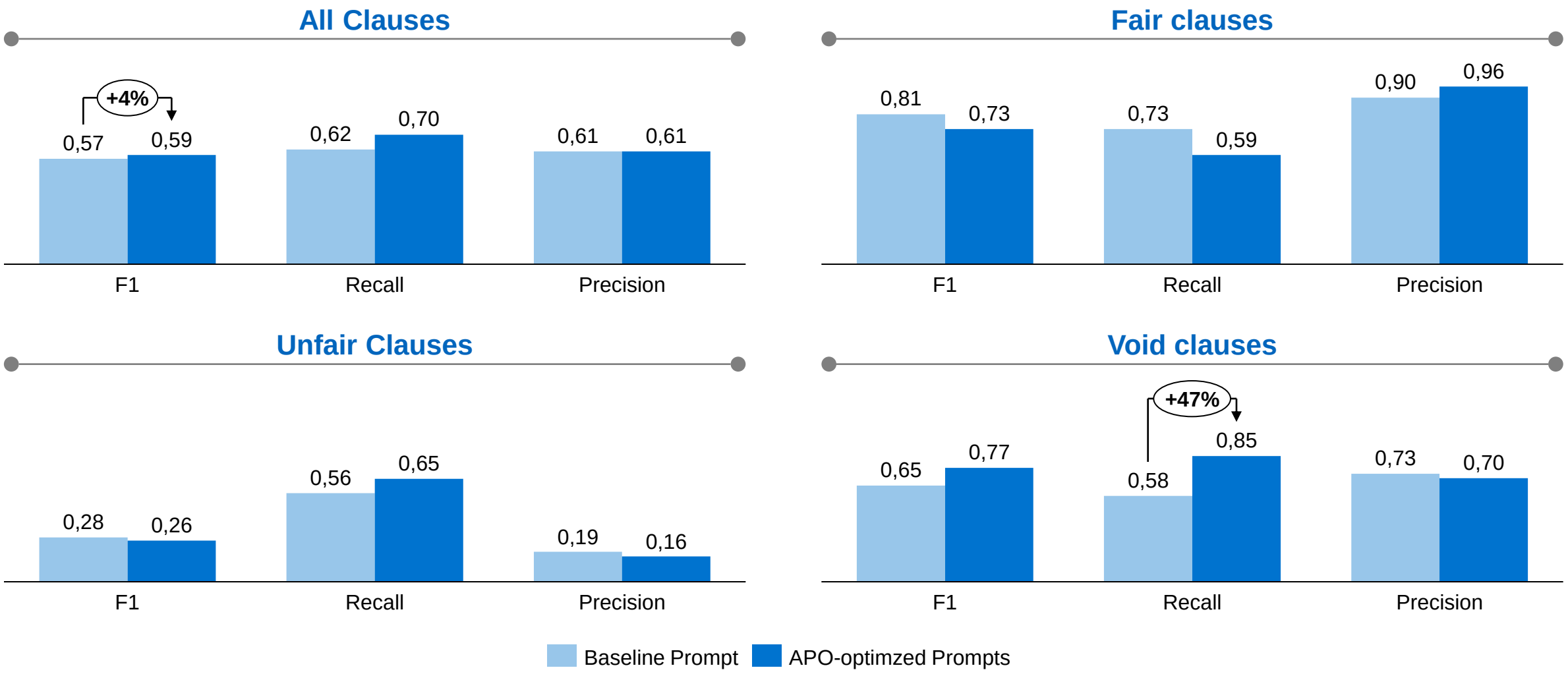
Results

Discussion & Outlook

© sebis

While overall F1 improvement fell short of expectation, recall improvement for void clauses by 47% can be seen as a success due to high FN costs

Final Evaluation Performance of APO-Optimzed Prompts relative to Baseline Prompt

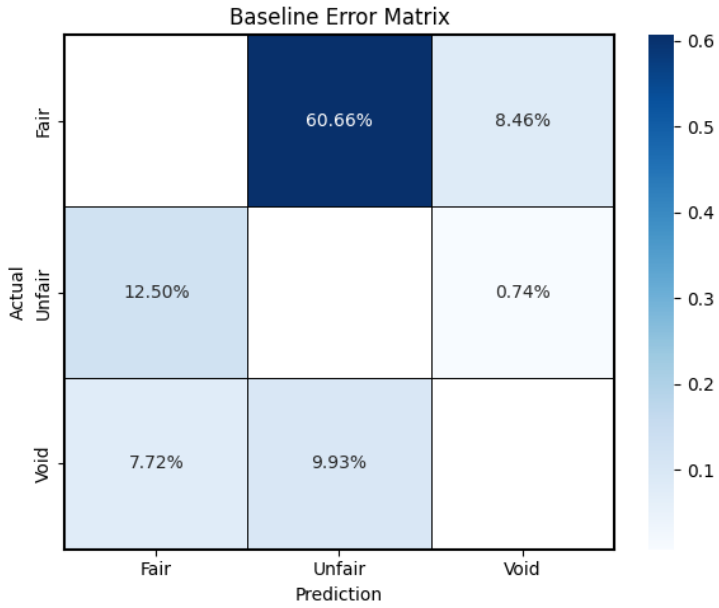


© sebis

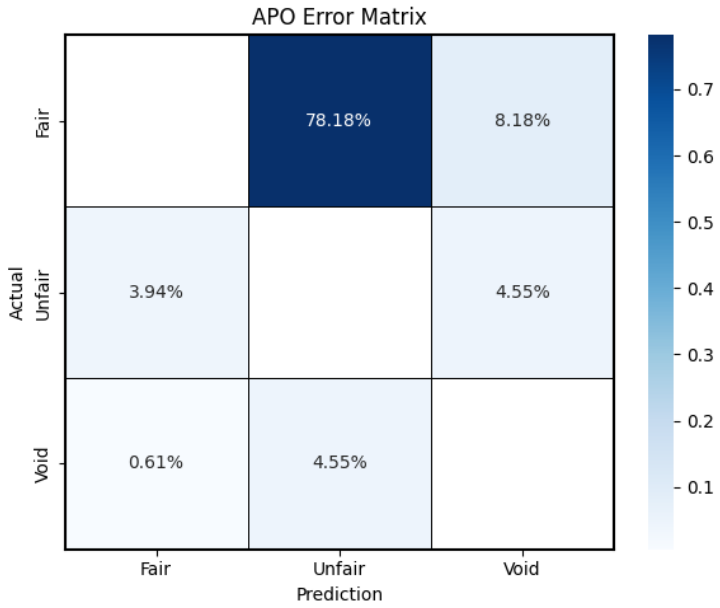
APO shifts classification towards a risk-averse strategy, prioritizing conservative decisions without necessarily improving fairness understanding

Error matrix produced by APO-optimized prompts relative to Baseline Prompt

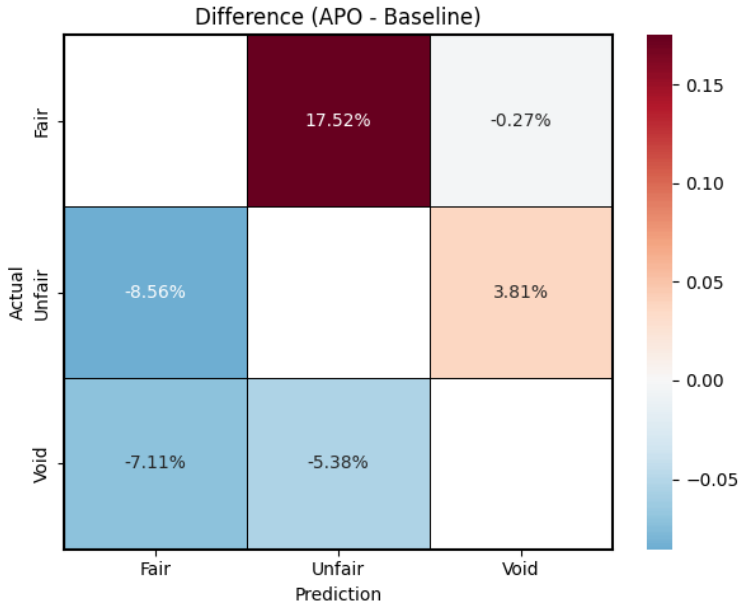
Baseline



APO



Delta



The LLM struggles to define the decision boundary between fair und unfair clauses and when in doubt defaults to the more conservative choice

© sebis

Outline

Background & Motivation

Problem Statement & Research Questions

Methodology

Results

Discussion & Outlook

© sebis

APO yielded mixed results: 47% recall gain for void clauses but only 4% F1 improvement, indicating potential for policy refinement yet limited gains

Main Findings & Implications

Main Takeaway

4%

of overall **F1-score improvement** of APO over baseline suggest that APO alone may not be the most effective approach for improving clause classification

Promising Results

47%

increase in **recall for void clauses** demonstrates APO's capability to enhance detection of void clauses, offering potential mechanism for **automated policy refinement**

Implications for KIBeKodA

Future research should explore alternative methods to improve performance, integrating APO within **hybrid approach**:

- Few-Shot Learning
- Retrieval-augmented generation
- (Parameter-Efficient) Fine-Tuning

© sebis

Our findings must be viewed in the light of some limitations relating to data size & quality, limited knowledge in legal domain and single researcher bias

Limitations & Future Work

1

Limited Data Size & Inter-Annotater Agreement

Clause corpus extension; Synthetic clause generation; Explore continuous fairness scale

2

Legal Domain Knowledge Gap

Collaborative check with lawyers to ensure alignment of derived policies with existing legal frameworks and statutory requirements

3

Single Researcher Bias

Feedback loop with practicing lawyers to check validity of APO-derived prompts

© sebis

Prof. Dr.
Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

We use an LLM-based optimization setup in which the LLM is used for all stages of prompt optimization process

Setup of simplified Demo

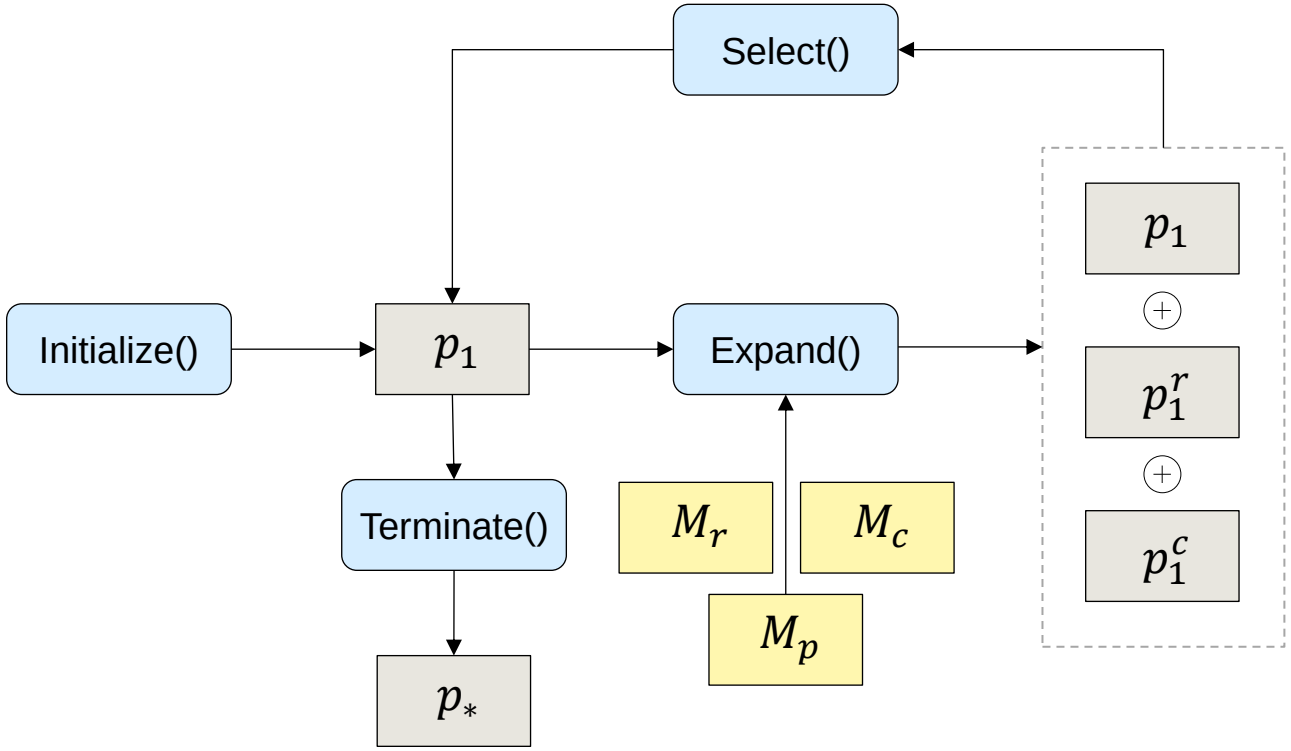
30 Clauses

Category **Termination**

9 void clauses, **3 unfair** clauses, **18 fair** clauses

No initial polices provided by lawyers

Simplified APO Process



© sebis

Meta-Prompt Component Refinement

Du bist ein Prompt Engineer, spezialisiert darauf, Prompt-Formulierungen für einen LLM-Klassifizierer zu verbessern. Der Klassifizierer ist darauf spezialisiert, Klauseln der Kategorie {Kategorie} aus Arbeitsverträgen hinsichtlich ihrer rechtlichen Zulässigkeit und Fairness zu bewerten und sie in einer der folgenden Kategorien zu klassifizieren: 'unbedenklich', 'potentiell unzulässig', 'unzulässig'.

In der letzten Iteration wurde der Klassifizierer-Prompt verwendet, um mehrere Klauseln zu klassifizieren. Einige dieser Klauseln wurden jedoch falsch klassifiziert. Für jede falsch klassifizierte Klausel sind folgende Informationen angegeben:

- Die tatsächliche Bewertung durch Anwälte, die als „ground truth“ gilt (als „Bewertung“ aufgeführt),
- Die vom LLM vorgenommene Klassifikation (als „Klassifikation“ aufgeführt) und
- Die Begründung des LLMs für seine Klassifikation (als „Erklärung LLM“ aufgeführt).
- Eine optionale Begründung der Anwälte, warum die Klausel der „ground truth“ entspricht (aufgeführt als „Erklärung Anwalt“, dieses Feld kann leer sein).

Inputdaten:

- **Klauseln**: Alle Klauseln, die in der letzten Iteration vom LLM falsch klassifiziert wurden, befinden sich unter ### Klauseln.
- **Aktuelle Klassifierungskomponente**: Der aktuelle Prompt, der vom Klassifizierer verwendet wurde, ist ebenfalls bereitgestellt. Dieser Prompt beinhaltet die „Klassifizierungskomponente“, welche beschreibt, unter welchen Bedingungen eine Klausel als 'unzulässig' und wann als 'potentiell unzulässig' einzustufen ist.
- **Verteilung**: Tatsächliche Verteilung der Klauseln aus der Kategorie ###Kategorie### in die Bewertungskategorien 'unbedenklich', 'potentiell unzulässig', 'unzulässig'. Die Verteilung ist als wahr anzunehmen, weil sie von professionellen Anwälten kommt.

Deine Aufgabe:

Erstelle eine verbesserte Version der „Klassifizierungskomponente“, die den Klassifizierer so anleitet, dass die falsch klassifizierten Klauseln bei der nächsten Iteration korrekt eingestuft werden. Die neue Komponente sollte:

1. Die spezifischen Bedingungen für die Klassifikationen einer Klausel als 'unbedenklich' und 'potentiell unzulässig' so präzisieren, dass alle unter ### Klauseln gelisteten Fehler vermieden werden. Unter welchen Bedingungen eine Klausel als unzulässig zu klassifizieren ist, soll in der Komponente nicht erklärt werden. Ein besonderes Augenmerk soll nicht auf den eindeutigen, sondern auf den Grenzfällen liegen. Berücksichtige dabei auch die Erklärung des LLM-Klassifizierers für seine Bewertung.
2. Inkorporiere auch die Verteilung der Klauseln in die einzelnen Bewertungskategorien bei der Erstellung, allerdings nur indikativ, ohne die konkrete Verteilung in Zahlen zu nennen.
3. Verfasse die neue Klassifizierungskomponente ohne unnötige Einleitungs- und Schlussformulierungen. Die Komponente soll direkt beginnen mit den Kriterien und keine zusätzlichen Titel oder Erläuterungssätze wie „Die Klassifizierungskomponente wird wie folgt definiert“ enthalten.

Nutze das folgende Format, um die neue Komponente abzugrenzen:
<START>...Das ist ein Platzhalter für die neue Komponente...<ENDE>

Klauseln:
{Klauseln}

Aktuelle Klassifizierungskomponente:
{Aktuell Komponente}

Verteilung:
{Verteilung}

Passe die Komponente so an, dass alle bereitgestellten, falsch klassifizierten Klauseln in der nächsten Iteration vom Klassifizierer richtig bewertet werden.

© sebis

Meta-Prompt Policy Refinement

```
Du bist ein Prompt Engineer, spezialisiert darauf, Prompt-Formulierungen für einen LLM-Klassifizierer zu verbessern. Der Klassifizierer ist darauf
spezialisiert, Klauseln der Kategorie {Kategorie} aus Arbeitsverträgen hinsichtlich ihrer rechtlichen Zulässigkeit und Fairness zu bewerten und sie in einer
der folgenden Kategorien zu klassifizieren: 'unbedenklich', 'potentiell unzulässig', 'unzulässig'.

Der Prompt besteht aus mehreren Teilen, einschließlich einer Liste von Regeln. Jede Regel beschreibt, unter welchen Bedingungen eine Klausel als 'unzulässig'
einzustufen ist.

Aktuell kommt es allerdings zu zwei Fehlermodi:
1. Fehlermodus 1: Der LLM-Klassifizierer erkennt Klauseln, die von Anwälten als unzulässig bewertet wurden, nicht als unzulässig.
2. Fehlermodus 2: Der LLM-Klassifizierer klassifiziert Klauseln als unzulässig, obwohl diese von Anwälten nicht als unzulässig erachtet werden.

Für jede Klausel, auf die entweder Fehlermodus 1 oder Fehlermodus 2 zutrifft, sind folgende Informationen angegeben:
- Die tatsächliche Bewertung durch Anwälte, die als „ground truth“ gilt (als „Bewertung“ aufgeführt).
- Die vom LLM vorgenommene Klassifikation (als „Klassifikation“ aufgeführt).
- Die Begründung des LLMs für seine Klassifikation (als „Erklärung LLM“ aufgeführt).
- Eine optionale Begründung der Anwälte, warum die Klausel der „ground truth“ entspricht (aufgeführt als „Erklärung Anwalt“, dieses Feld kann leer sein).
- Eine Liste von IDs der Regeln, die das LLM angewendet hat, um die Klausel zu klassifizieren (als Policy List“ aufgeführt).

### Klauseln, die falsch klassifiziert wurden:
{Klauseln}

### Aktuelle Regelmenge:
{Policies}

Deine Aufgabe ist es, die Menge der Regeln so anzupassen, dass das LLM beim nächsten Durchlauf in der Lage ist, alle Klauseln mithilfe der Gesamtheit der
Regeln korrekt zu klassifizieren, sodass weder Fehlermodus 1 noch 2 vorkommt. Stelle sicher, dass die Regelmenge nach Hinzufügen der neuen Regeln keine
inhaltlichen Redunanzen aufweist.

Zum Beispiel:
1) <START>Add() -> Eine Klausel, die in den AGB ein generelles Pfändungs- oder Abtretungsverbot für die Forderungen des Arbeitnehmers festlegt, ist nach §
308 Nr. 9 lit. a BGB unwirksam und damit unzulässig<ENDE>
2) <START>Modify(352fgdjsdgd) -> Eine Klausel, die in den AGB ein generelles Pfändungs- oder Abtretungsverbot für die Forderungen des Arbeitnehmers festlegt,
ist nach § 308 Nr. 9 lit. a BGB unwirksam und damit unzulässig<ENDE>
```


© sebis

Meta-Prompt Crossover Candidates

Du bist ein Prompt Engineer, spezialisiert darauf, Prompt-Formulierungen für einen LLM-Klassifizierer zu verbessern. Der Klassifizierer ist darauf spezialisiert, Klauseln der Kategorie {Kategorie} aus Arbeitsverträgen hinsichtlich ihrer rechtlichen Zulässigkeit und Fairness zu bewerten und sie in einer der folgenden Kategorien zu klassifizieren: 'unbedenklich', 'potentiell unzulässig', 'unzulässig'.

Vorgaben:
1. **Klassifizierungskomponente A** und **Klassifizierungskomponente B**: Diese beiden Komponenten bieten jeweils eine spezifische Grundlage für die Klassifikation und definieren die Bedingungen, unter denen eine Klausel als 'unbedenklich' und 'potentiell unzulässig' gilt. Beide Komponenten können unterschiedliche Schwerpunkte und Formulierungen aufweisen, jedoch sollen ihre besten Elemente in der neuen Komponente zusammengeführt werden.

2. **Kombinationsstrategie**:
- Wähle die klarsten und präzisesten Anweisungen aus jeder Komponente und vereine diese so, dass eine umfassende, leicht verständliche Anleitung entsteht.
- Berücksichtige dabei, dass die neue Komponente die Kernprinzipien beider Originalkomponenten beibehält, aber in einer optimierten, strukturierten und widerspruchsfreien Form formuliert ist.
- Vermeide Redundanzen und streiche sich überschneidende oder widersprüchliche Inhalte.

Klassifizierungskomponente A
{Komponente A}

Klassifizierungskomponente B
{Komponente B}

Aufgabe:
Erstelle die neue Klassifizierungskomponente aus Klassifizierungskomponente A und B und grenze sie mit den folgenden Tags ab. Verfasse die neue Klassifizierungskomponente ohne Einleitungsformulierungen. Die Komponente soll direkt beginnen und keine zusätzlichen Titel oder Erläuterungssätze wie „Die Klassifizierungskomponente wird wie folgt definiert“ enthalten.

<START>...Das ist ein Platzhalter für die neue Komponente...<ENDE>

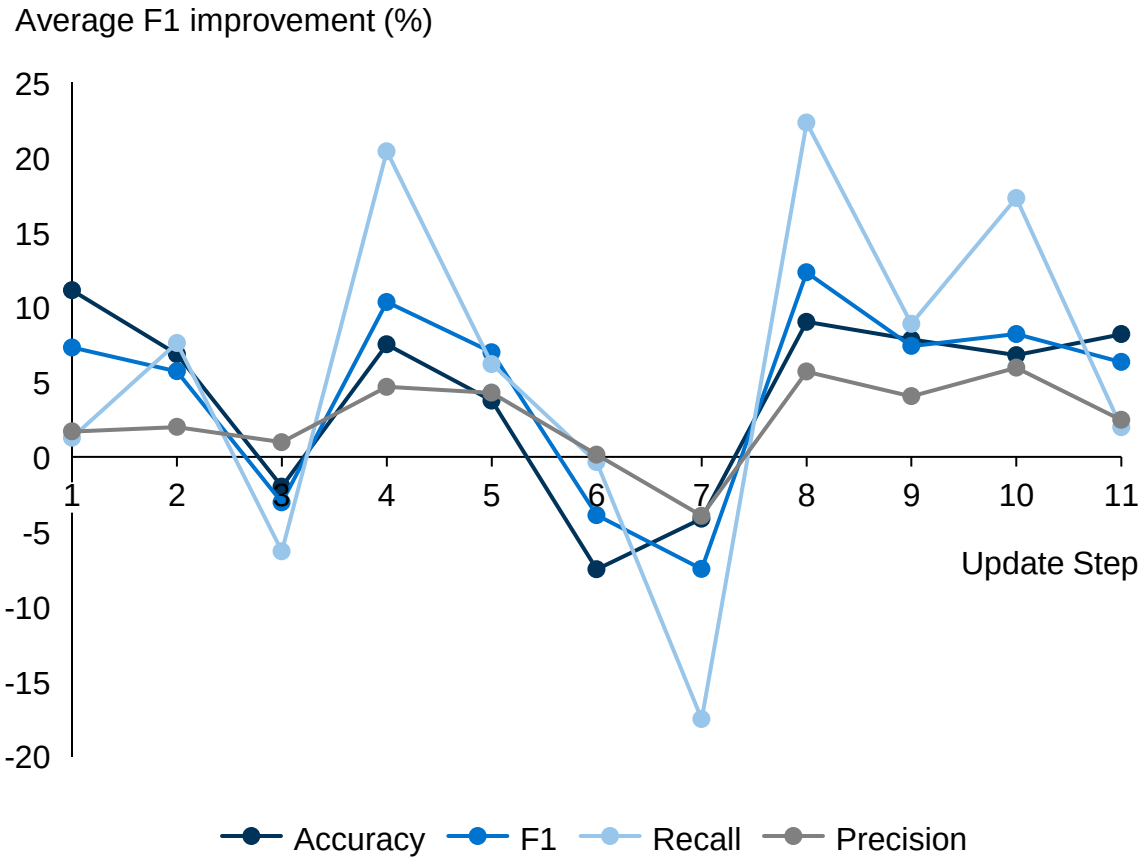
Die neue Komponente soll die Kriterien aus beiden Klassifizierungskomponenten optimal vereinen, um die Genauigkeit der Klassifikation in zukünftigen Iterationen zu maximieren.

© sebis

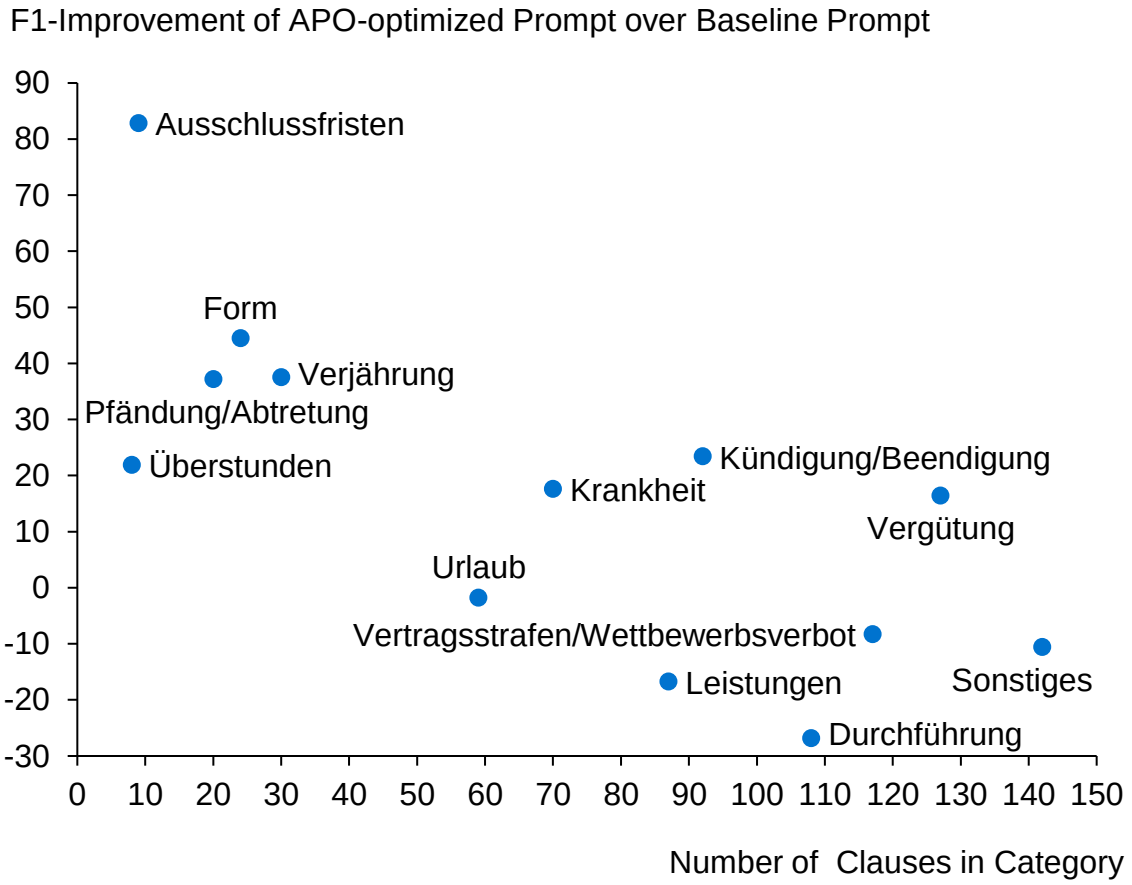
Relatively stable optimization trajectory together with negative correlation between category clause size and F1 improvement suggest overfitting

Final Evaluation Performance of APO-Optimzed Prompts relative to Baseline Prompt

Optimization Trajectory



Final Performance by Category



© sebis

Two macro-level hypothesis emerge that might explain why APO did not achieve its expected improvement potential

Main Finding

State of the art APO methods achieved only a F1 performance increase of 4% over an expert written baseline prompt for fairness classification of clauses out of german employment contracts

Algorithmic & Model Limitations

- Expert-Designed vs. Randomized Prompt Initialization
- Instruction-Level vs. Knowledge-Level Optmization

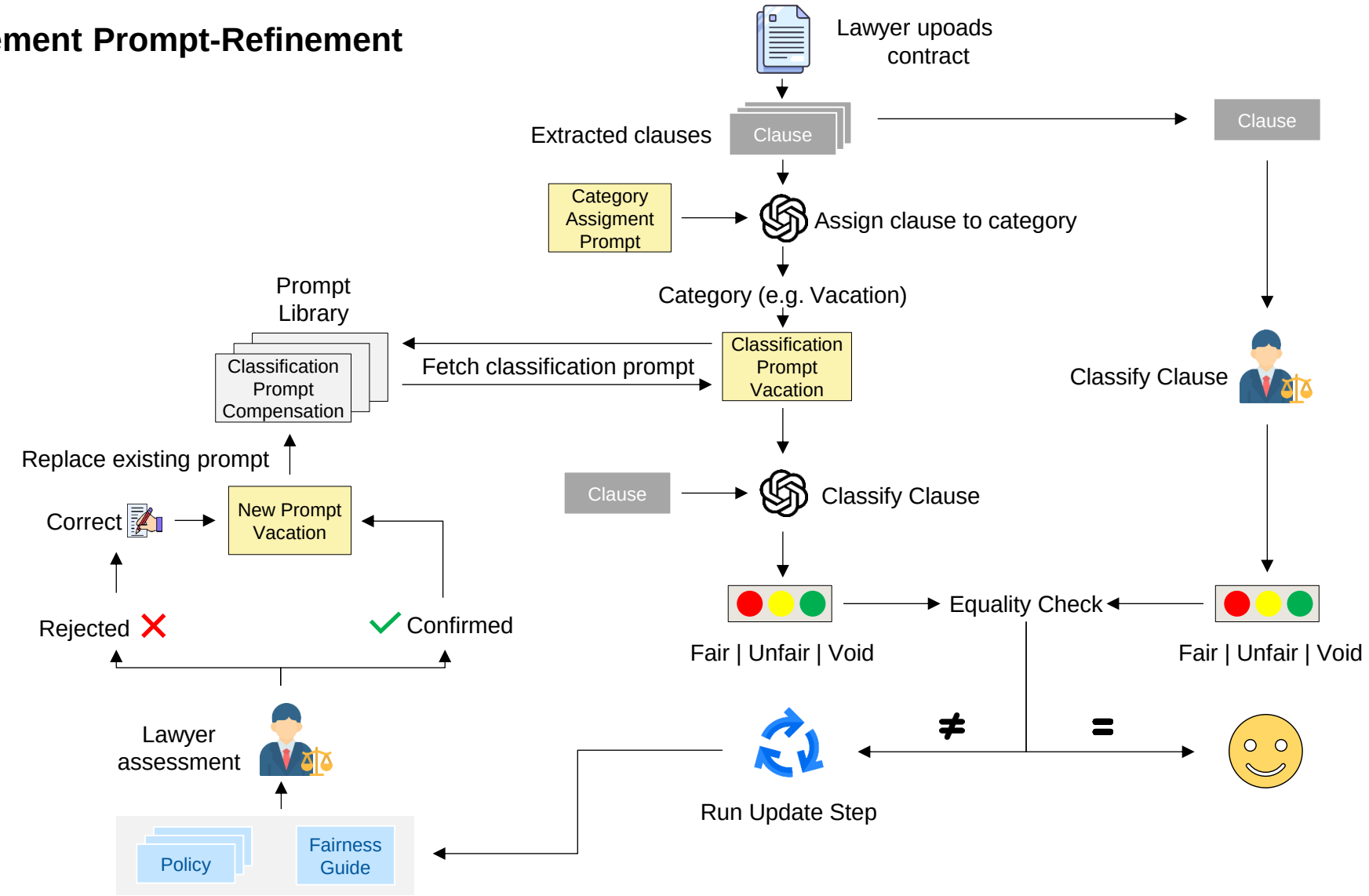
Data Limitations

- Quantitative Deficiencies
- Qualitative Deficiencies

© sebis

47% improvement in void clause detection enables APO-driven self-improvement of polcies in KIBEKodA

Self-Improvement Prompt-Refinement



© sebis

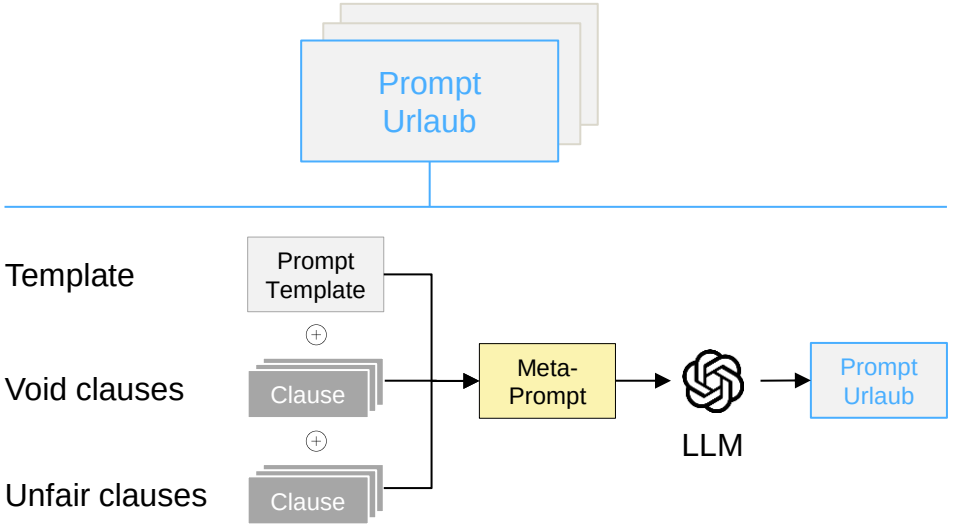
Redesigned Prompt Methodology

1 → 2

Manual Design of Prompt Template

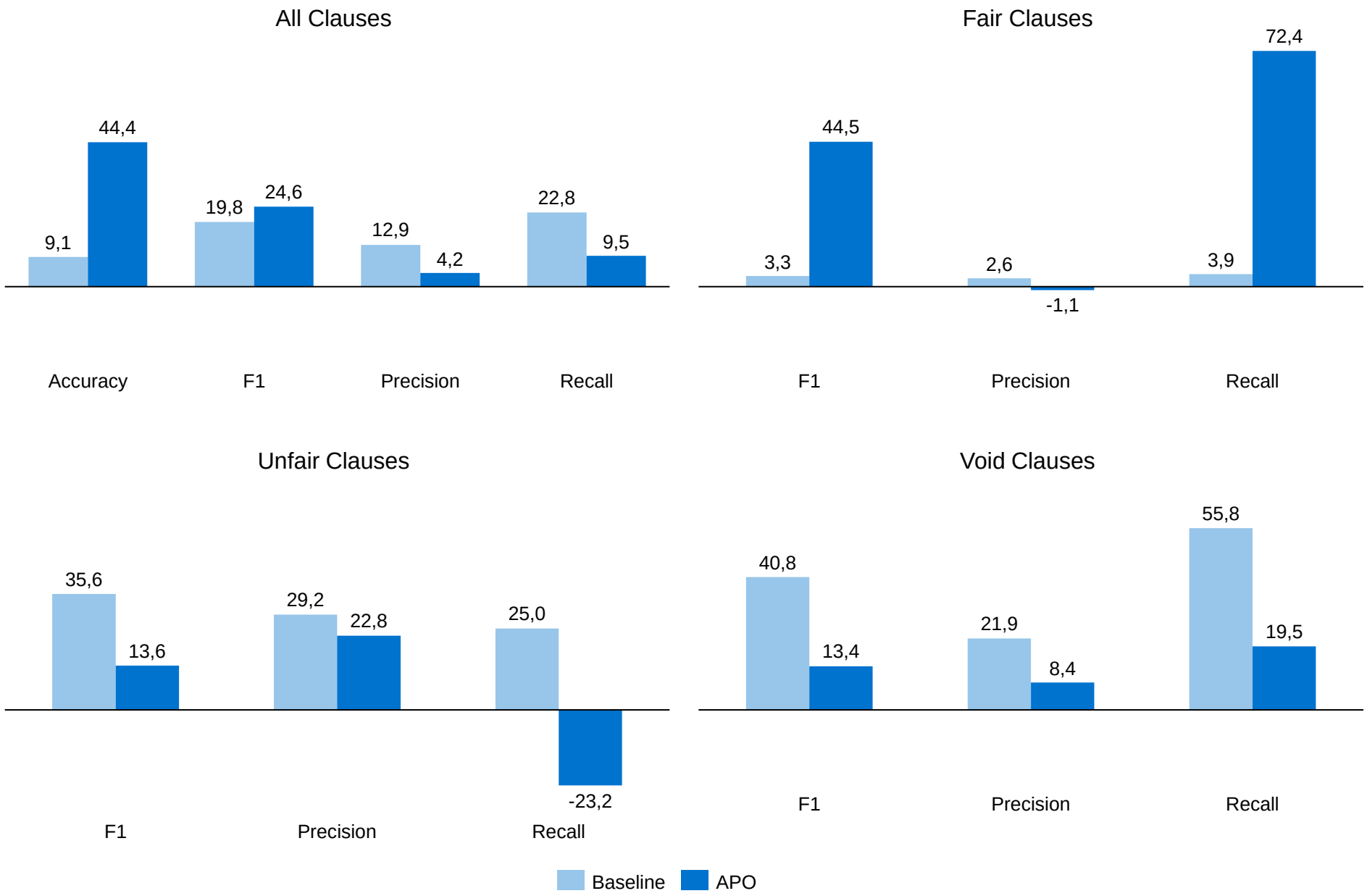
```
### Aufgabe ###
Bewerte die folgende Klausel als unbedenklich, …, unfair.

### Anweisungen ###
1. Eine Klausel ist unzulässig, wenn:
-   <Regel 1 für unzulässig>

2. Eine Klausel ist unzulässig, wenn:
-   <Regel 1 für unzulässig>

3. Falls keine Regel zutrifft, bewerte als unbedenklich

### Klausel ###
{Klausel}
```

Automatic Meta-Prompt Guided Creation of Prompts



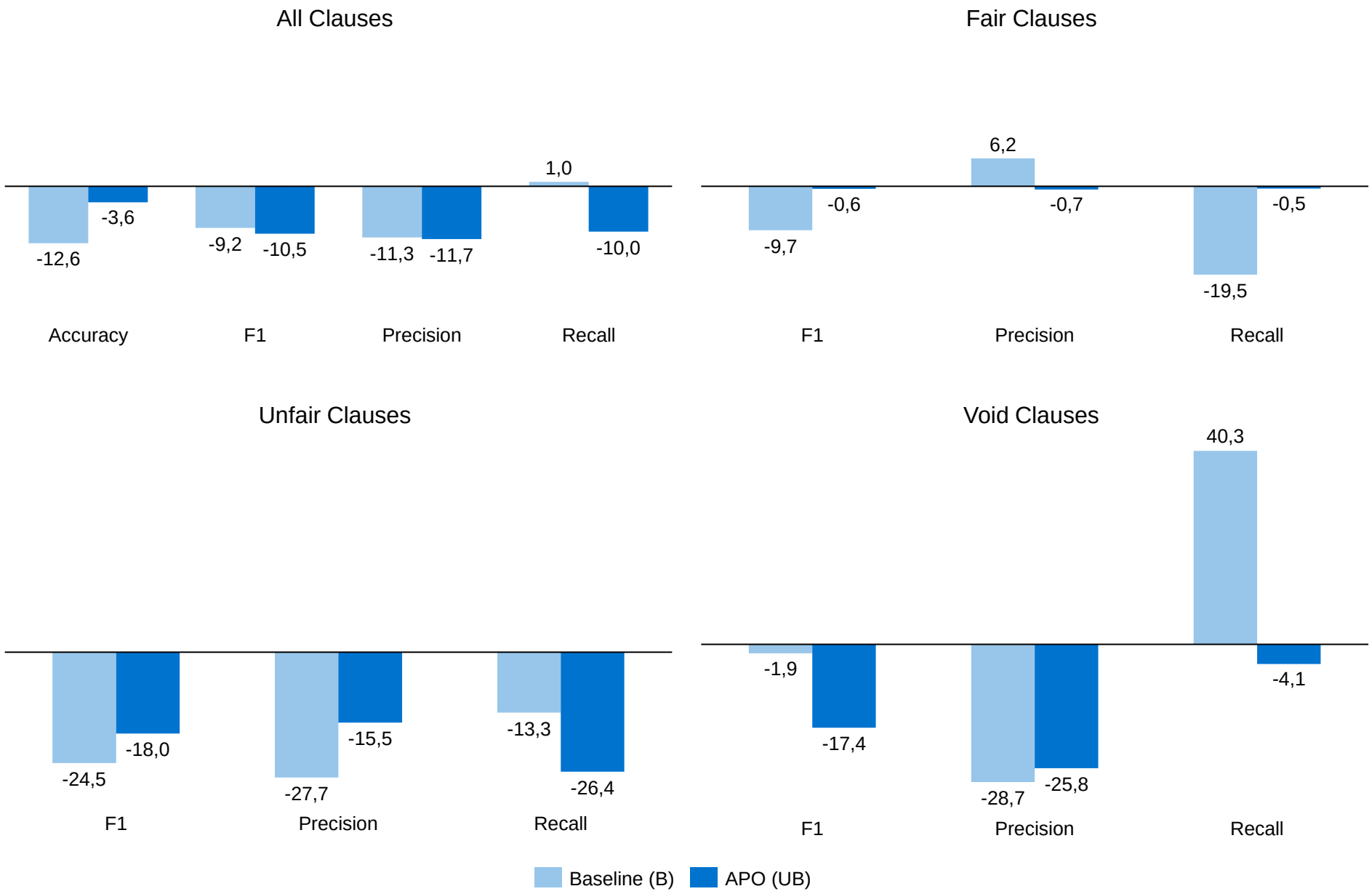
© sebis

Relative Impact by using GPT-4o over GPT-4o-mini (in %)



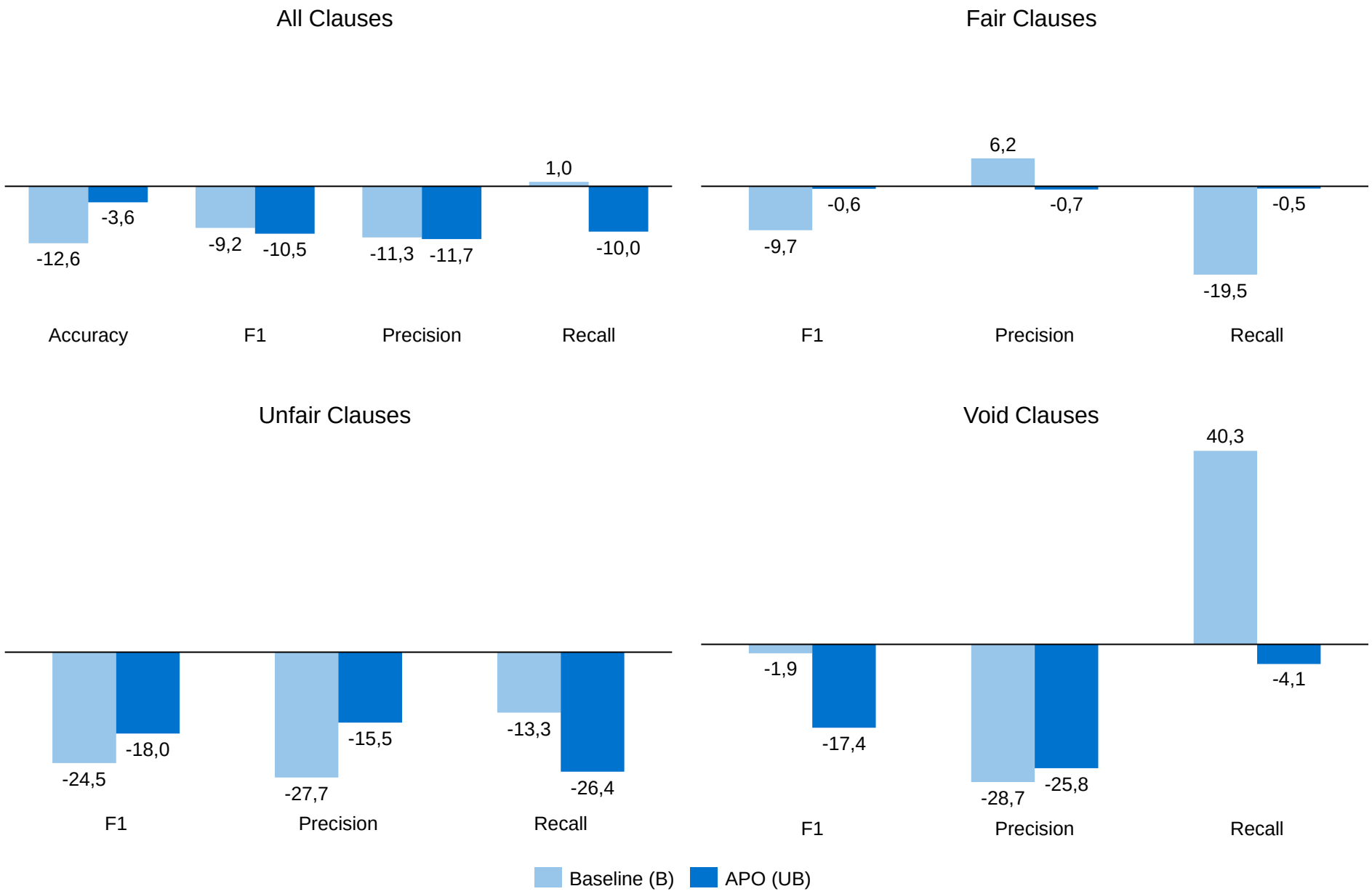
© sebis

Relative Improvement of Static Prompts (in %)



© sebis

Relative Improvement of Static Prompts (in %)



© sebis

Configuration (APO, GPT-4o)													
Category	Overall				Fair			Unfair			Void		
	Accuracy	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
Other	0.69	0.56	0.60	0.70	0.81	0.70	0.96	0.21	0.60	0.13	0.67	0.50	1.00
Compensation	0.72	0.64	0.75	0.68	0.80	0.68	0.97	0.27	0.75	0.17	0.86	0.82	0.90
Penalty Clause	0.50	0.51	0.63	0.60	0.57	0.40	1.00	0.30	0.79	0.18	0.67	0.71	0.62
Execution	0.50	0.44	0.68	0.57	0.62	0.45	0.98	0.27	0.91	0.16	0.00	0.00	0.00
Termination	0.68	0.44	0.54	0.43	0.83	0.74	0.94	0.07	0.12	0.05	0.44	0.75	0.30
Tasks	0.49	0.46	0.71	0.51	0.62	0.46	0.97	0.09	0.67	0.05	0.67	1.00	0.50
Illness	0.80	0.67	0.76	0.66	0.89	0.81	1.00	0.30	0.60	0.20	0.82	0.88	0.78
Vacation	0.59	0.60	0.75	0.60	0.68	0.54	0.93	0.38	0.70	0.26	0.75	1.00	0.60
Statute of Limitations	0.62	0.56	0.62	0.60	0.58	0.44	0.88	0.40	0.50	0.33	0.71	0.92	0.58
Format	0.58	0.57	0.63	0.64	0.62	0.45	1.00	0.43	0.43	0.43	0.67	1.00	0.50
Garnishment	0.85	0.72	0.88	0.75	0.77	0.62	1.00	0.40	1.00	0.25	1.00	1.00	1.00
Overtime	0.62	0.69	0.83	0.75	0.67	0.50	1.00	0.40	1.00	0.25	1.00	1.00	1.00
Cut-off Periods	0.89	0.80	0.94	0.75	0.00	0.00	0.00	0.67	1.00	0.50	0.93	0.88	1.00
Overall	0.63	0.59	0.7	0.61	0.73	0.59	0.96	0.26	0.65	0.16	0.77	0.85	0.70

© sebis

Configuration (Baseline, GPT-4o)													
Category	Overall				Fair			Unfair			Void		
	Accuracy	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
Other	0.74	0.63	0.78	0.59	0.84	0.75	0.96	0.25	0.60	0.16	0.80	1.00	0.67
Compensation	0.69	0.55	0.56	0.58	0.79	0.73	0.86	0.12	0.25	0.08	0.75	0.71	0.80
Penalty Clause	0.63	0.56	0.60	0.62	0.75	0.67	0.85	0.34	0.64	0.23	0.59	0.48	0.77
Execution	0.75	0.61	0.74	0.61	0.84	0.75	0.96	0.37	0.73	0.25	0.00	0.00	0.00
Termination	0.67	0.36	0.45	0.36	0,79	0,71	0.89	0.29	0.62	0.19	0.00	0.00	0.00
Tasks	0.68	0.55	0.71	0.57	0.80	0.67	0.98	0.13	0.67	0.07	0.73	0.80	0.67
Illness	0.74	0.57	0.69	0.56	0.86	0.77	0.98	0.38	0.8	0.25	0.47	0.50	0.44
Vacation	0.80	0.62	0.57	0.77	0.89	0.89	0.89	0.45	0.50	0.42	0.50	0.33	1.00
Statute of Limitations	0.53	0.41	0.39	0.45	0.59	0.50	0.73	0.00	0.00	0.00	0.64	0,67	0,62
Format	0.54	0.40	0,45	0.42	0.65	0.91	0.50	0.55	0.43	0.75	0.00	0.00	0..0
Garnishment	0.65	0.53	0.47	0.62	0.88	0.88	0.88	0.00	0.00	0.00	0.71	0.55	1.00
Overtime	0.50	0.52	0.78	0.58	0.50	0.33	1.00	0.40	1.00	0.25	0.67	1.00	0.50
Cut-off Periods	0.78	0.44	0.44	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.88	0.88
Overall	0.70	0.58	0.62	0.61	0.81	0.73	0.90	0.28	0.56	0.19	0.65	0.58	0.73

© sebis

Configuration (APO, GPT-4o-mini)													
Category	Overall				Fair			Unfair			Void		
	Accuracy	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
Other	0.47	0.4	0.48	0.61	0.61	0.44	0.98	0.2	0.9	0.11	0.4	0.5	0.33
Compensation	0.57	0.55	0.63	0.74	0.63	0.47	0.95	0.26	1.0	0.15	0.76	0.74	0.78
Penalty Clause	0.3	0.33	0.53	0.49	0.29	0.17	0.93	0.26	0.93	0.15	0.43	0.38	0.5
Execution	0.48	0.43	0.57	0.67	0.6	0.43	0.98	0.26	0.91	0.15	0.0	0.0	0.0
Termination	0.33	0.36	0.47	0.55	0.43	0.28	0.96	0.14	0.62	0.08	0.5	0.75	0.38
Tasks	0.29	0.43	0.59	0.74	0.35	0.22	1.0	0.09	1.0	0.05	0.83	1.0	0.71
Illness	0.37	0.39	0.62	0.56	0.48	0.32	1.0	0.19	1.0	0.1	0.5	0.38	0.75
Vacation	0.42	0.38	0.49	0.52	0.49	0.33	1.0	0.36	0.9	0.22	0.29	0.33	0.25
Statute of Limitations	0.57	0.42	0.53	0.43	0.55	0.38	1.0	0.0	0.0	0.0	0.71	0.92	0.58
Format	0.5	0.46	0.66	0.6	0.17	0.09	1.0	0.63	0.71	0.56	0.6	1.0	0.43
Garnishment	0.7	0.55	0.71	0.75	0.4	0.25	1.0	0.25	1.0	0.14	1.0	1.0	1.0
Overtime	0.88	0.86	0.83	0.94	0.91	0.83	1.0	0.67	1.0	0.5	1.0	1.0	1.0
Cut-off Periods	0.78	0.44	0.44	0.44	0.0	0.0	0.0	0.0	0.0	0.0	0.88	0.88	0.88
Overall	0.44	0.47	0.58	0.64	0.51	0.34	0.98	0.23	0.85	0.13	0.68	0.71	0.65

© sebis

Configuration (Baseline, GPT-4o-mini)													
Category	Overall				Fair			Unfair			Void		
	Accuracy	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
Other	0.7	0.34	0.36	0.41	0.82	0.72	0.95	0.19	0.5	0.12	0.0	0.0	0.0
Compensation	0.65	0.51	0.6	0.53	0.8	0.74	0.86	0.14	0.38	0.09	0.6	0.47	0.84
Penalty Clause	0.51	0.48	0.56	0.58	0.62	0.49	0.87	0.33	0.86	0.21	0.47	0.38	0.62
Execution	0.81	0.56	0.55	0.57	0.89	0.87	0.91	0.22	0.27	0.19	0.0	0.0	0.0
Termination	0.65	0.32	0.33	0.36	0.78	0.71	0.86	0.19	0.38	0.13	0.0	0.0	0.0
Tasks	0.69	0.5	0.54	0.55	0.82	0.71	0.97	0.07	0.33	0.04	0.6	0.6	0.6
Illness	0.61	0.42	0.46	0.51	0.79	0.67	0.97	0.21	0.6	0.12	0.27	0.25	0.29
Vacation	0.66	0.37	0.35	0.39	0.78	0.76	0.8	0.32	0.4	0.27	0.0	0.0	0.0
Statute of Limitations	0.33	0.31	0.42	0.38	0.52	0.38	0.86	0.12	0.5	0.07	0.29	0.25	0.33
Format	0.46	0.21	0.15	0.33	0.63	1.0	0.46	0.0	0.0	0.0	0.0	0.0	0.0
Garnishment	0.4	0.31	0.39	0.27	0.33	0.25	0.5	0.0	0.0	0.0	0.6	0.55	0.67
Overtime	0.62	0.4	0.42	0.56	0.8	0.67	1.0	0.4	1.0	0.25	0.0	0.0	0.0
Cut-off Periods	0.56	0.36	0.42	0.31	0.0	0.0	0.0	0.0	0.0	0.0	0.71	0.62	0.83
Overall	0.64	0.48	0.54	0.51	0.78	0.7	0.88	0.21	0.44	0.14	0.46	0.37	0.6

© sebis