ArgAssist: LLM-based Argument Synthesis for Insurance Disputes

Anubhav Sinha, Nitin Ramrakhiyani, Sachin Pawar, Isha Narang, Manoj Apte

The 27th Meeting of the Special Interest Group on Discourse and Dialogue

Motivation

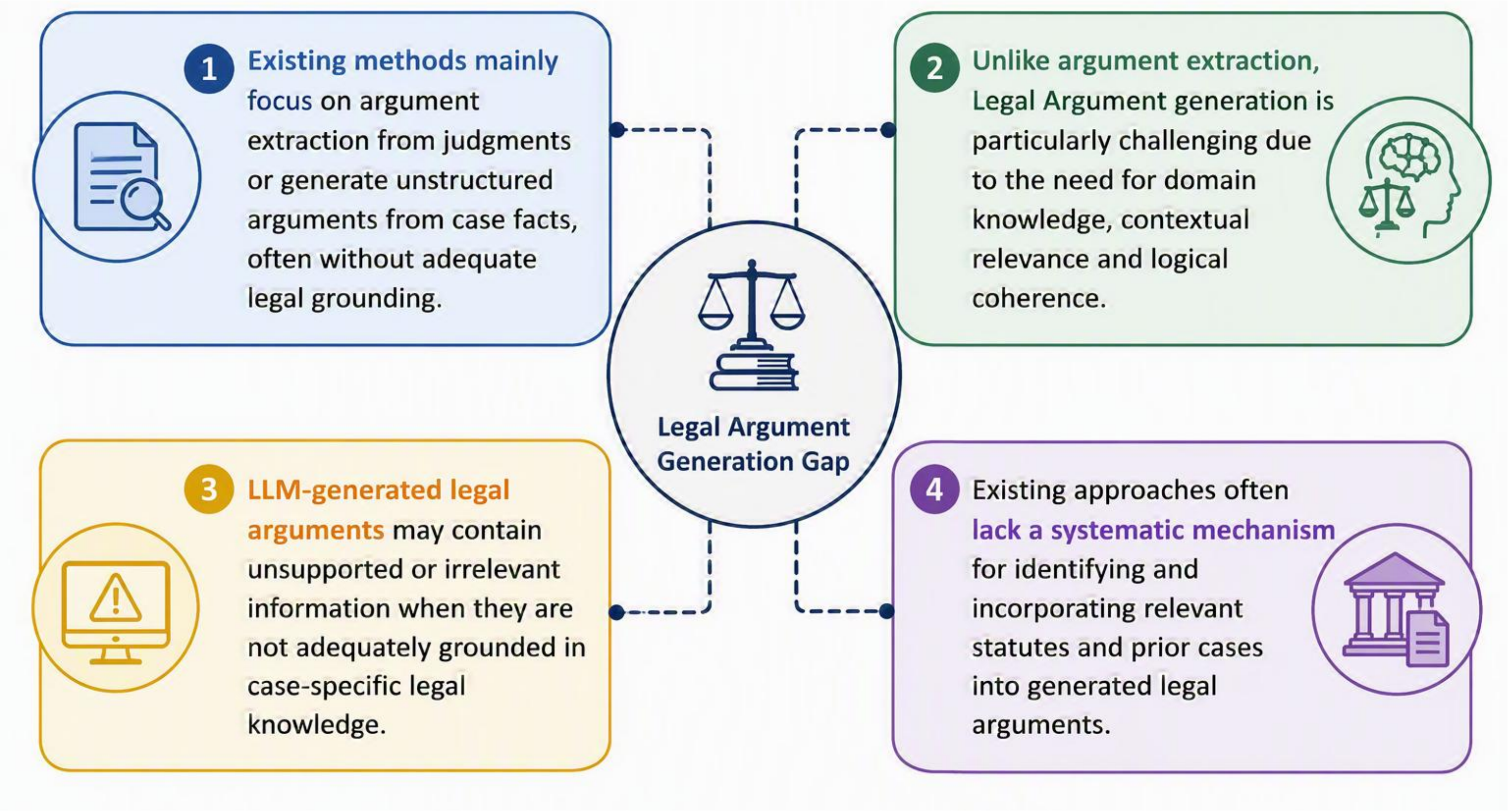
- **Legal disputes are frequent and complex**, while access to timely and affordable legal assistance remains limited.

- Most of these disputes end up in litigation, **straining the already overburdened judicial systems.**

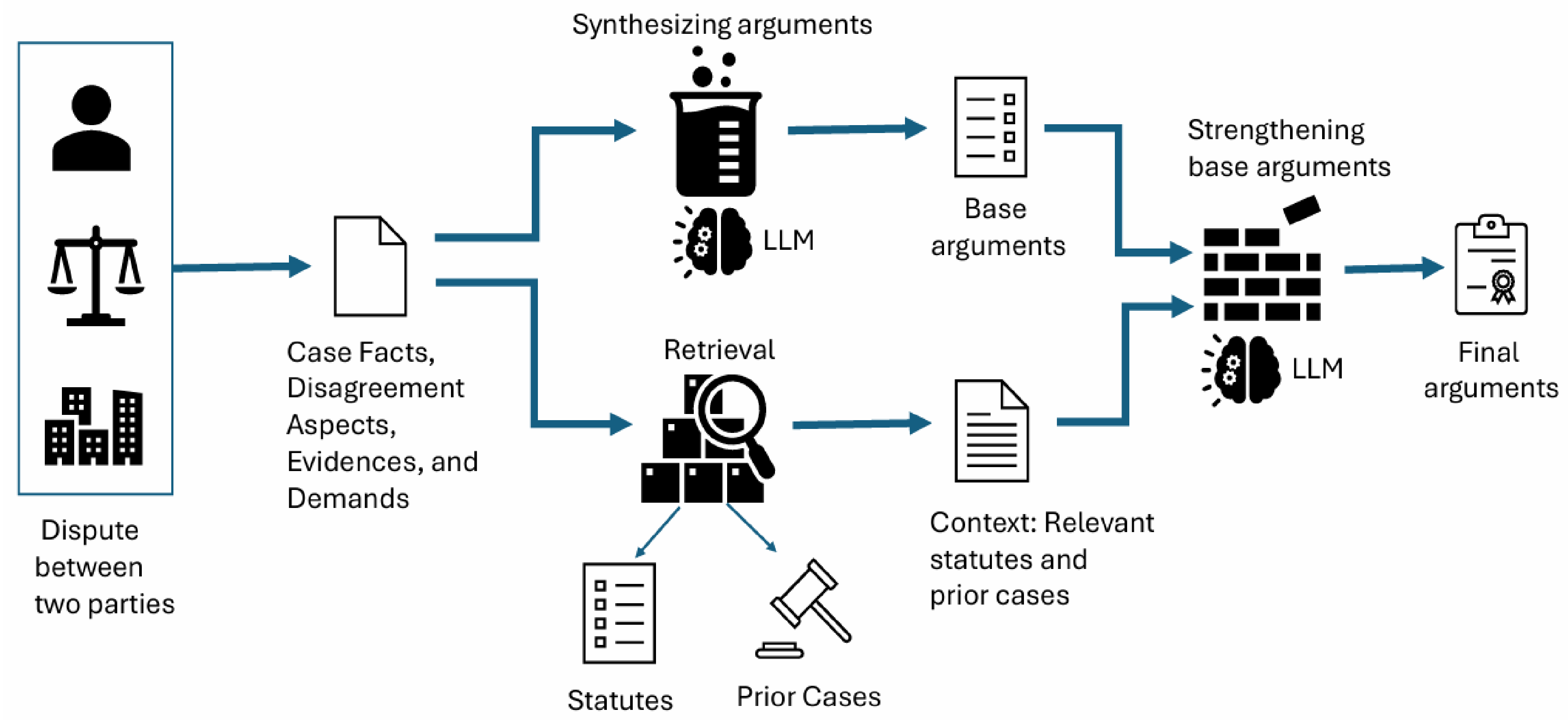
- There is a **need to democratize legal reasoning and support individuals** who may lack access to professional legal representation.

- This motivates the **need for automated legal reasoning systems** capable of generating legally grounded arguments for dispute resolution.

Problem statement & Research gap



ArgAssist Pipeline



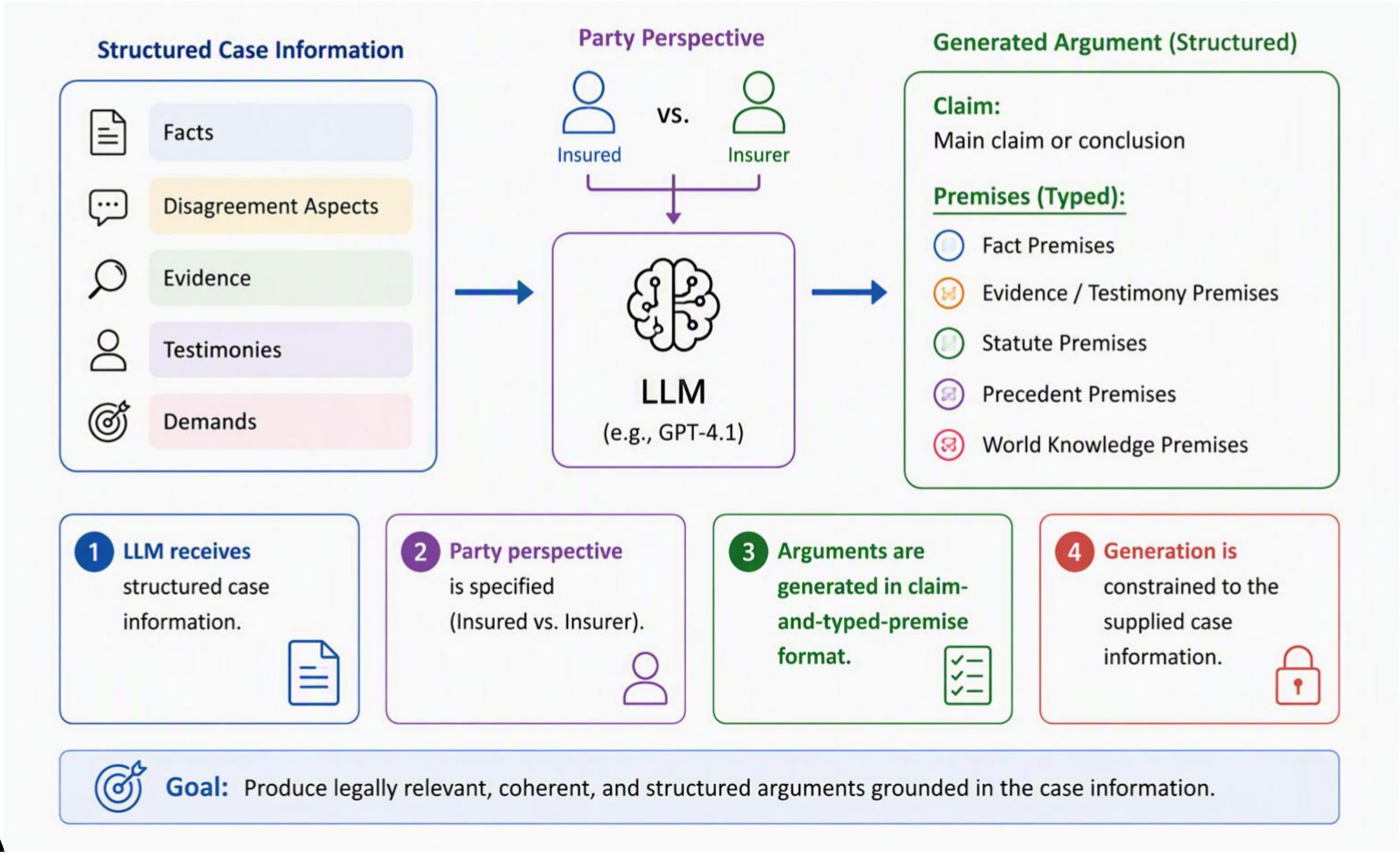
ArgAssist

- **Multi-Stage Argument Synthesis:**
  - Generates party-specific legal arguments from structured dispute information.
- **Structured Argument Representation:**
  - Organizes arguments as claims supported by multiple evidence types.
- **Dispute-Aware Legal Knowledge Retrieval:**
  - Identifies relevant statutes and prior cases using dispute-specific information.
- **Retrieval-Augmented Refinement:**
  - Strengthens arguments by incorporating legally relevant authorities.

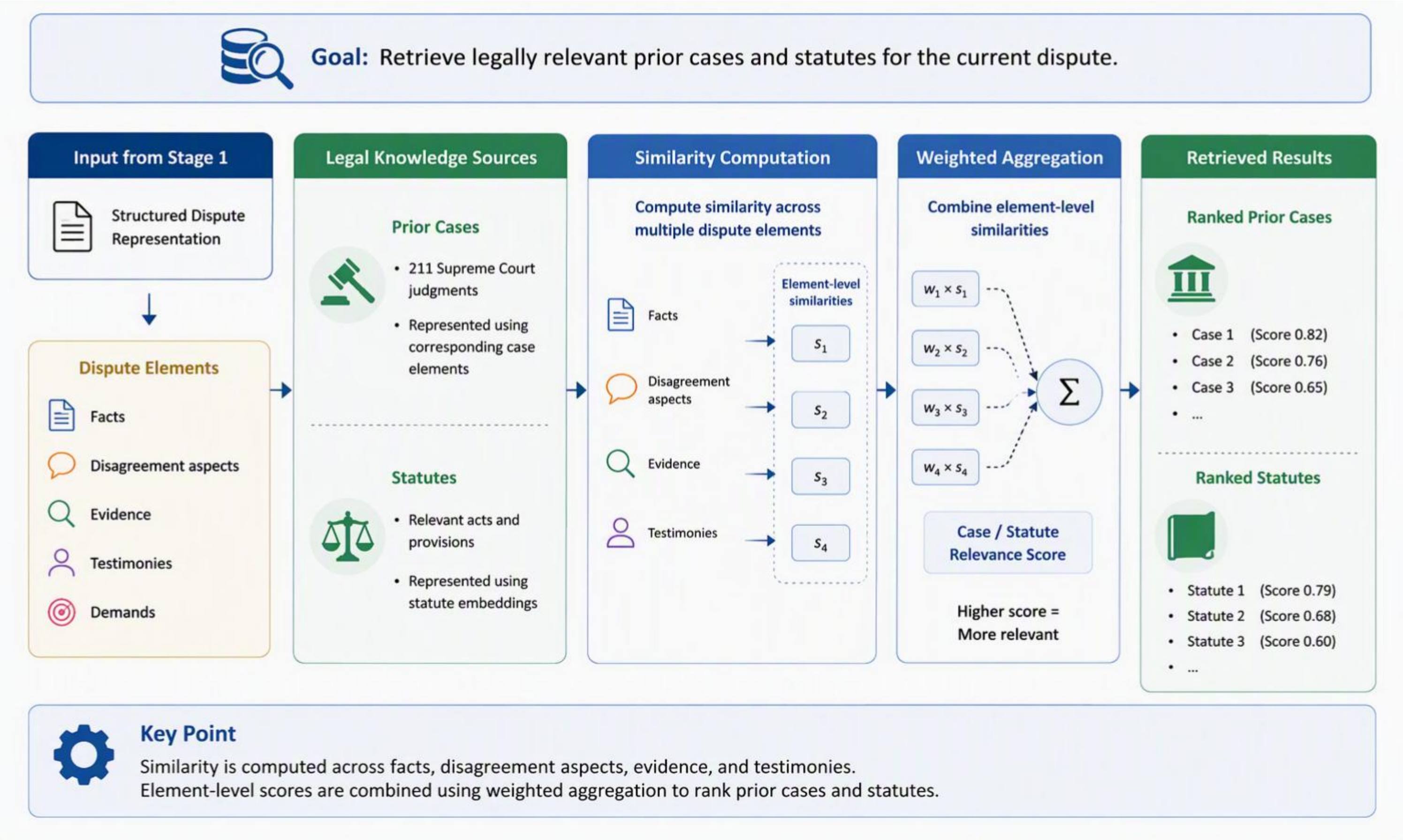
Argument Representation

- **Claim-centric representation:** Every argument revolves around a clear claim or legal position.
- **Multi-source support:** Claims are supported using facts, evidence, statutes, precedents, and domain knowledge.
- **Improved traceability:** Each supporting premise can be linked back to its underlying source information.
- **Better legal grounding:** Explicit representation of legal authorities reduces hallucination and strengthens argument quality.

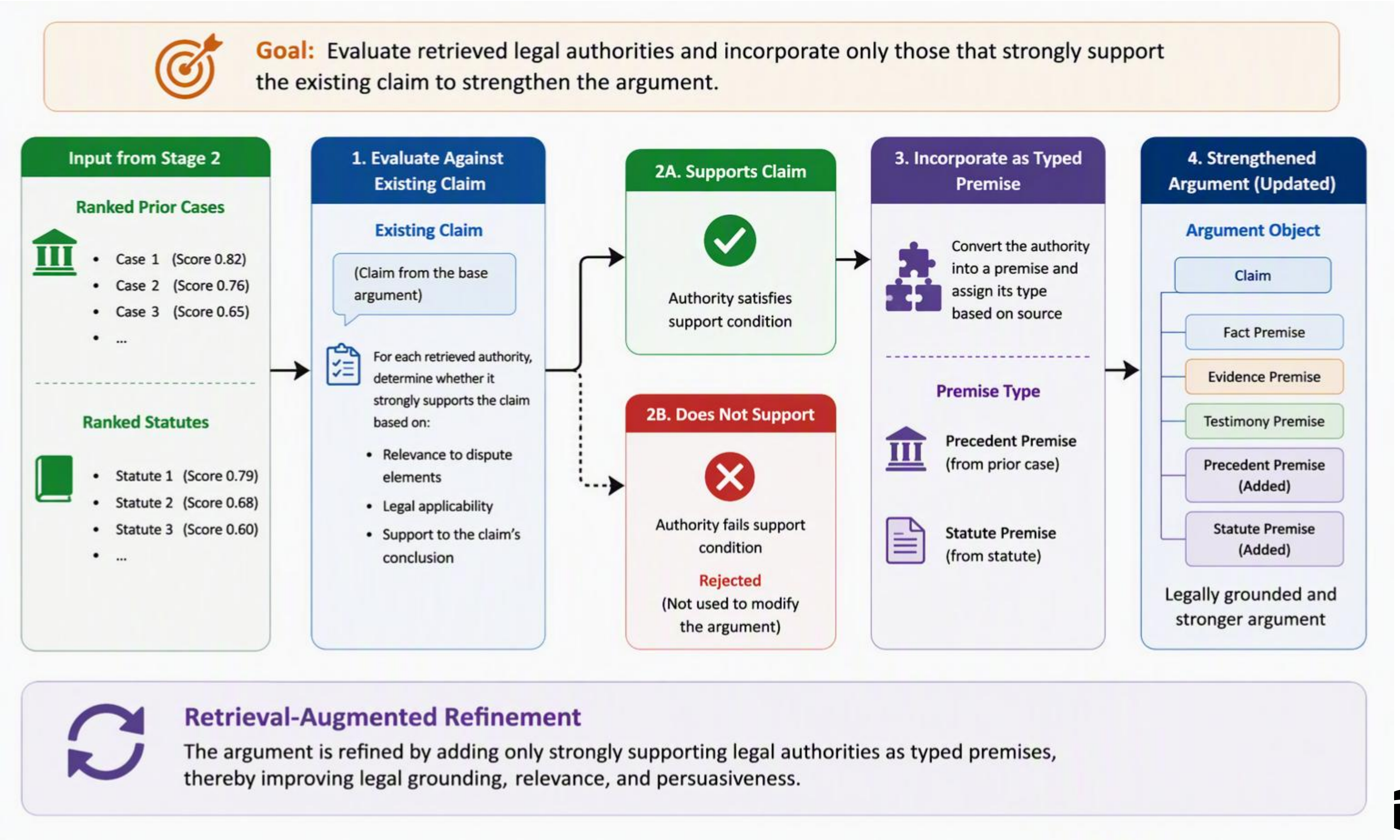
Stage 1, Base Argument Synthesis



Stage 2, Retrieval of Legal Knowledge



Stage 3, Argument Strengthening



Dataset and experimental setup

Dataset	*Cases*
VID	*129*
LID	*82*
Prior cases	*211*

- **Experimental Configuration:** 10 disputes from each dataset were used for development, while the remaining disputes formed the test sets. Experiments used GPT-4.1 and text-embedding-3-large.

- **Baseline and Evaluation:** ArgAssist was compared with a Flan-T5-based argument generation baseline and evaluated using Retrieval Quality, the proposed ArgMatch metric, and LLM-based quality assessment of unmatched arguments.

Evaluation methodology

ArgMatch (AM)

- Compares generated arguments with gold-standard court arguments.
- Uses claim-level bipartite matching and premise-level semantic similarity.
- Reports Precision, Recall, and F1 scores.

Quality Assessment of Unmatched Arguments (QSFP)

- Evaluates generated arguments that do not have corresponding gold-standard matches.
- Uses an LLM-as-a-judge.
- Assesses logical coherence, factual accuracy, legal foundation, and persuasiveness.

Result and key findings

Consistent best performance: ArgAssist_final achieves the highest AM scores across both datasets and both parties.

Stronger after alignment: Adding relevant statutes and precedents consistently improves ArgAssist_base

Promising court argument coverage: AM recall improves significantly for the strengthened system.

High-quality additional arguments: Unmatched arguments are of high quality and score 3.51-3.97 out of 5.

Dataset	Party	Technique	Micro-averaged			Macro-averaged			QS_{FP}
			AM_P	AM_R	AM_{F1}	AM_P	AM_R	AM_{F1}	
VID	insured	baseline	0.076	0.097	0.086	0.082	0.119	0.092	1.441
		ArgAssist_base	0.271	0.524	0.357	0.275	0.525	0.341	3.155
		ArgAssist_final	**0.288**	**0.558**	**0.380**	**0.293**	**0.568**	**0.365**	**3.622**
	insurance company	baseline	0.152	0.196	0.171	0.178	0.241	0.191	1.445
		ArgAssist_base	0.332	0.627	0.434	0.332	0.629	0.418	3.168
		ArgAssist_final	**0.343**	**0.648**	**0.449**	**0.342**	**0.652**	**0.431**	**3.513**
LID	insured	baseline	0.080	0.113	0.094	0.084	0.128	0.094	1.701
		ArgAssist_base	0.204	0.479	0.286	0.209	0.517	0.286	3.438
		ArgAssist_final	**0.227**	**0.533**	**0.319**	**0.232**	**0.581**	**0.318**	**3.977**
	insurance company	baseline	0.085	0.134	0.104	0.103	0.190	0.124	1.660
		ArgAssist_base	0.241	0.551	0.335	0.243	0.587	0.333	3.472
		ArgAssist_final	**0.260**	**0.595**	**0.362**	**0.262**	**0.638**	**0.360**	**3.771**

AM indicates the proposed ArgMatch metric and QSFP indicates LLM-as-a-judge scores assigned for unmatched generated arguments

Conclusion and future work

- **Effective Legal Argument Synthesis:** ArgAssist generates structured, dispute-specific, and legally grounded arguments for both parties involved in an insurance dispute.
- **Benefits of Legal Knowledge Refinement:** Strengthening base arguments with relevant statutes and prior cases consistently improves alignment with arguments presented in actual court proceedings.
- **Extension to Other Domains:** Future work will explore disputes in areas such as finance, e-commerce, consumer goods, and taxation.
- **Broader Evaluation and Knowledge Coverage:** The framework will be evaluated using larger prior-case collections, additional legal knowledge such as legal factors, and a wider range of proprietary and open-source LLMs.

Thank you