POLITECNICO DI TORINO
Repository ISTITUZIONALE

Trustworthy Regulatory Requirements Extraction using Local LLM: A Multi-Agent vs. Single-Prompt Benchmark on the EU AI Act and GDPR

Original
Trustworthy Regulatory Requirements Extraction using Local LLM: A Multi-Agent vs. Single-Prompt Benchmark on the EU AI Act and GDPR / Ahmed, M., Naz, M.A.. - ELETTRONICO. - (2026), pp. 1-5. (2026 IEEE 2nd International Conference on AI and Emerging Technology for Sustainable Future Catania (Ita) 24-25 July 2026) [10.1109/ICAISF70757.2026.11767456].

Availability:
This version is available at: 11583/3015996 since: 2026-09-29T11:43:04Z

Publisher:
IEEE

Published
DOI:10.1109/ICAISF70757.2026.11767456

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

29 November 2026

Trustworthy Regulatory Requirements Extraction using Local LLM: A Multi-Agent vs. Single-Prompt Benchmark on the EU AI Act and GDPR

Mishaal Ahmed
Department of Mathematics & Physics,
University of Campania,
Caserta, Italy
mishaal.ahmed@studenti.unicampania.it

Muhammad Ajmal Naz
Department of Electronics & Telecom.,
Politecnic di Torino,
Turin, Italy
muhammadajmal.naz@studenti.polito.it

***Abstract*—Regulated AI systems need requirements that link testable SHALL statements to statutory source text, yet local large language models often invent obligations or misstate modality when elicitation is under-specified. We ask whether a sequential Analyst–Engineer–QA multi-agent workflow (B5) outperforms fair baselines when every role shares one Llama 3 8B model via Ollama on EU AI Act and GDPR chunks. Using quote-validated expert gold, the structured single-prompt baseline B1 achieves the highest F1 on the EU AI Act test (0.382) and GDPR generalization (0.337), ahead of B5 (0.320, 0.322) and unstructured B0 (0.269, 0.275). B5 exhibits higher hallucination than B1 on both splits (e.g., 0.134 vs 0.009 on EU test) despite three LLM calls per chunk versus one for B1. Ablations B0–B5 indicate that precision gains, not added agents, drive quality on consumer hardware. We answer the primary research question in the negative for matched prompt content: role-separated inference (B5) does not beat the single-call structured prompt (B1) on F1. For trustworthy, privacy-preserving regulatory requirements engineering, practitioners should prefer single-call structured prompts and deploy multi-agent pipelines only when intermediate artifacts are required for human audit.**

***Index Terms*—Trustworthy AI, Requirements Engineering, Regulatory Compliance, Multi-agent LLM, EU AI Act, GDPR, Sustainable Governance.**

I. Introduction

Deploying AI in regulated domains requires traceable links from software requirements to statutory text—a prerequisite for responsible innovation aligned with societal expectations and instruments such as the EU AI Act [1] and GDPR [2]. Organizations must show that system obligations derive from authentic legal sources, not from model confabulation, before those requirements enter design documents, conformity assessments, or data-protection impact analyses. Requirements engineering (RE) therefore faces a dual challenge: accelerate elicitation from lengthy consolidations, yet preserve auditability—every SHALL statement should be defensible against a cited span in the source regulation.

Large language models (LLMs) promise to bridge natural-language statutes and structured requirements, but unconstrained prompting often *invents* obligations, conflates permissions with duties, or omits exceptions buried in sub-articles [3]. These failures are especially costly in compliance settings, where a plausible but unsupported requirement can propagate through downstream testing and governance workflows. Recent multi-agent designs decompose elicitation into analyst, author, and verifier roles, orchestrated sequentially through frameworks such as CrewAI [4]. The intuition is that specialization reduces error; however, when every role shares one small model on consumer hardware, additional inference steps may compound mistakes rather than add independent capacity [5].

Parallel to fidelity concerns, many public-sector and industry adopters prefer *local* inference—via runtimes such as Ollama [6]—to limit data egress, control cost, and support sustainable deployment on modest GPUs [7]. Yet benchmarks for regulatory RE rarely combine (i) authentic held-out statutory splits, (ii) quote-level grounding checks on each predicted requirement, (iii) fair baselines that match multi-agent instructions in a single prompt [8], and (iv) explicit measurement of inference cost per chunk. Without such controls, claims that "more agents yield better compliance requirements" remain difficult to verify. While broader frameworks exist for legal NLP [9] and generalized agent evaluation [10], without these specific controls, claims that "more agents yield better compliance requirements" remain difficult to verify.

We address this gap with a controlled study on English EUR-Lex consolidations of the EU AI Act (primary test corpus) and GDPR Articles 25–32 (generalization). All conditions share one Llama 3 8B backbone at temperature 0. We compare zero-shot elicitation (B0), a structured single-call prompt that encodes the same role instructions as the full pipeline (B1), partial two-agent ablations (B2–B4), and a three-agent Analyst→Engineer→QA workflow (B5). Outputs are evaluated against quote-validated expert reference requirements using precision, recall, F1, and hallucination rate (predictions whose cited spans are absent from the source chunk).

RQ: Under matched prompts and fixed model capacity, does the three-agent pipeline (B5) beat structured single-prompt elicitation (B1) on requirement-level F1 against quote-

validated expert gold?

Contributions:

- A reproducible regulatory RE workflow with frozen prompts and ablations B0–B5, isolating orchestration depth while holding model weights constant.
- Dual-corpus evaluation (EU AI Act test; GDPR generalization) with requirement-level metrics and systematic hallucination measurement on gold-annotated chunks.
- An empirical negative result: B1 achieves the highest F1 on both corpora; B5 beats unstructured B0 but not B1, despite three LLM calls per chunk versus one for B1.
- Actionable guidance for trustworthy local RE: prefer single-call structured prompts on 8B-class models; deploy multi-agent pipelines when human reviewers need intermediate constraints and drafts, not when F1 alone is the objective.

II. Related Work

a) LLM-based requirements engineering.: NLP for requirements engineering has progressed from template-based extraction to neural and LLM-based paraphrasing of natural-language specifications into structured statements [5]. Recent systems prompt general-purpose models to emit EARS-style requirements or to iterate over drafts with critic modules. Multi-agent frameworks assign personas such as analyst, author, and verifier, often orchestrated sequentially through toolchains such as CrewAI [4]. However, many evaluations compare a multi-step pipeline only against a weak zero-shot baseline, without a *matched* single prompt that encodes the same instructions in one call.

b) Regulatory text and compliance NLP.: Legal and regulatory NLP addresses citation structure, modality (obligations vs. permissions), and span-level entailment between statutes and downstream artifacts [1], [2]. The EU AI Act and GDPR create demand for traceable links from software requirements to source articles, yet published LLM studies rarely release held-out statutory splits [11], [12], quote-level grounding checks, and systematic hallucination metrics on requirement tuples [13]. Our work targets that gap with authentic EUR-Lex consolidations and a reproducible quote-validation protocol.

c) Grounding, local deployment, and agent depth.: Retrieval-augmented and self-critique pipelines aim to reduce unsupported generations [14], but when every role shares one small model, additional calls may compound errors rather than add capacity [5]. Privacy-preserving and cost-sensitive deployments increasingly rely on local runtimes (e.g., Ollama [6]) rather than cloud APIs [15]. We contribute a controlled comparison on a fixed Llama 3 8B backbone: full ablation B0–B5, explicit LLM-call cost, and dual-corpus evaluation. Our finding that structured single-call B1 outperforms three-agent B5 on F1, while B5 still beats unstructured B0, offers actionable guidance for trustworthy regulatory RE on consumer hardware without overstating multi-agent benefits.

III. Methodology

Figure 1 summarizes the end-to-end system: statutory corpora are chunked and split; the primary B5 workflow invokes three role-separated agents through a CrewAI sequential orchestrator on a shared local Llama 3 model; verified requirements are scored against quote-validated expert gold (baselines B0–B1 follow the same data and evaluation path).

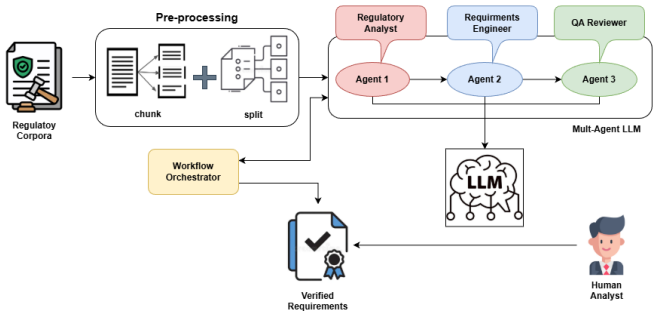

Fig. 1. System overview: regulatory corpora are preprocessed into chunks; B5 runs three sequential agent roles via CrewAI on a shared Ollama Llama 3 8B endpoint; outputs are evaluated against expert gold.

A. Pipeline and baselines

Our primary workflow (B5) maps each statutory chunk to verified requirements through three sequential CrewAI tasks on a shared local Llama 3 model at temperature 0. Agents do not converse directly: the orchestrator runs tasks in fixed order, and each later role receives earlier outputs as read-only context. The Regulatory Analyst extracts explicit constraints (actors, actions, conditions, scope); the Requirements Engineer formalizes them as *The system SHALL ... WHEN ...* statements; the QA Reviewer filters drafts against the source chunk and emits a verified requirement list with supporting quotes.

We compare B5 to five baselines (Table I) designed to isolate workflow structure and inference cost while holding model weights constant. B0 is an unstructured zero-shot lower bound: one call that asks for SHALL requirements without role decomposition. B1 is the fair matched comparator for our RQ: one call whose prompt encodes the same Analyse–Formalize–Verify instructions as B5 in a single message. B2–B4 reuse the same frozen role prompts in partial or reordered pipelines: B2 runs Analyst→Engineer (two calls, no QA); B3 runs Engineer→QA on the raw chunk, ablating the Analyst (two calls); B4 permutes role order as a sanity baseline (Engineer→Analyst→QA, three calls). All prompt templates are frozen (version 1.0) and held constant across conditions so that score differences reflect orchestration depth, not ad hoc retuning.

TABLE I
Baseline conditions (Llama 3 8B)

ID	Workflow	Calls
B0	Zero-shot SHALL extraction	1
B1	Matched single prompt (same roles as B5)	1
B2	Analyst → Engineer	2
B3	Engineer → QA (no Analyst)	2
B4	Engineer → Analyst → QA (shuffled)	3
B5	Analyst → Engineer → QA (primary)	3

B. Expert gold and metrics

Expert reference requirements are produced through an LLM-assisted annotation pipeline with deterministic substring quote validation (every supporting span must appear in the source chunk). This yields a structured benchmark, not multi-rater lawyer adjudication; the gold builder uses the same model family as the baselines (disclosed as a limitation). We report precision, recall, F1, and hallucination rate (fraction of predicted requirements whose quotes fail validation) on splits restricted to chunks with non-empty expert annotations.

IV. Experimental Setup

A. Platform and inference environment

Experiments were executed on a single Windows 10/11 consumer workstation with an NVIDIA GPU providing approximately 6 GB of VRAM—sufficient for quantized Llama 3 8B inference only when runs are serialized. All conditions use the Ollama runtime with the `llama3` model tag (8-billion-parameter instruct variant, locally quantized), **temperature** $= 0$, and non-streaming generation to maximize reproducibility. The CrewAI orchestrator invokes one shared model endpoint for every agent role; multi-agent baselines therefore differ only in prompt structure and call count, not in backbone weights. Each HTTP generation request was bounded by a two-minute timeout; failed chunks were retried according to the runner's default policy. Total wall-clock time for baseline inference over both corpora was on the order of four hours, reflecting sequential multi-call pipelines on consumer hardware.

B. Corpora and data partitions

Primary corpus: English consolidation of Regulation (EU) 2024/1689 (EU AI Act), chunked at article or sub-article boundaries after PDF extraction from EUR-Lex. **Generalization corpus:** GDPR Articles 25–32 (security of processing and related obligations), chunked similarly from Regulation (EU) 2016/679. Held-out *test* split for the EU AI Act comprises 98 statutory chunks; cross-regulation *generalization* split for GDPR comprises 128 chunks. Expert reference requirements were produced for a priority subset (100 chunks) using the same local model family with mandatory quote-in-source validation; reported metrics use only chunks in each split that contain filled gold labels (43 EU test; 52 GDPR).

C. Procedure

For each corpus split we ran every applicable baseline on every chunk in that split, passing identical source text and recording structured outputs (constraints, drafts, and verified SHALL requirements as applicable). EU AI Act test: full ablation B0–B5 (one, one, two, two, three, and three LLM calls per chunk respectively). GDPR generalization: B0, B1, and B5 only, to test whether the B1 versus B5 ordering survives a second regulation. Runs were executed sequentially per baseline (no parallel multi-GPU batching) to respect VRAM limits. After inference, each predicted requirement list was aligned to gold via requirement-level matching on normalized SHALL statements.

TABLE II
EU AI Act test (expert gold, gold-chunks-only)

Baseline	F1	Precision	Recall	Halluc.
B0	0.269	0.209	0.377	0.000
B1	0.382	0.387	0.377	0.009
B2	0.321	0.279	0.377	0.117
B3	0.270	0.210	0.377	0.000
B4	0.289	0.241	0.360	0.000
B5	0.320	0.289	0.360	0.134

D. Measured outcomes

Primary (requirement-level, expert gold): precision and recall of predicted versus gold requirements, their harmonic mean (F1), and **hallucination rate**—the fraction of predicted items whose cited source span is absent from the chunk (quote validation failure). **Secondary:** SHALL-template format pass rate (regex on *The system SHALL … WHEN …*) and error-taxonomy counts (e.g., invented modality, missed obligation) from manual rule-based coding of mismatches. **Cost proxy:** number of LLM calls per chunk (1 for B0/B1; 2 for B2/B3; 3 for B5), reported alongside F1 in trade-off figures. All metrics are computed on gold-chunks-only subsets so that empty legal passages do not deflate scores.

V. Results

This section reports requirement-level metrics against quote-validated expert gold on annotated chunks only (43 EU AI Act test; 52 GDPR generalization). B1 (single structured prompt) achieves the highest F1 on both corpora and outperforms the three-agent pipeline B5; B5 beats unstructured B0 but not B1 under matched prompts on Llama 3 8B.

A. EU AI Act test (primary corpus)

Table II lists precision, recall, F1, and hallucination rate for the full B0–B5 ladder. B1 achieves the highest F1 (0.382), 0.062 above B5 (0.320) and 0.113 above unstructured B0 (0.269). Recall is identical for B0–B3 (0.377), indicating that matched gold obligations are recovered at the same rate; differences arise from precision—B1 emits fewer spurious SHALL statements (0.387) than B5 (0.289) or B0 (0.209). B0 attains zero quote-level hallucination yet the lowest F1: every cited span appears in the chunk, but many predictions are semantically incorrect obligations. Intermediate ablations clarify the role of depth: B2 (Analyst+Engineer, two calls) reaches F1 0.321, comparable to B5, with lower hallucination (0.117 vs 0.134), so the QA stage does not reliably improve grounding on this model. B3 (Analyst+QA) and B4 (Engineer+QA) do not exceed B1, confirming that neither analysis nor formalization alone explains B1's advantage.

B. GDPR generalization

Table III tests transfer to GDPR Articles 25–32. The ranking B1>B5>B0 holds (F1 0.337, 0.322, 0.275), with B1 leading B5 by 0.015 F1. Recall remains similar for B0 and B1 (0.423), while B5 trades a small recall drop (0.415) for

TABLE III
GDPR GENERALIZATION (EXPERT GOLD)

Baseline	F1	Precision	Recall	Halluc.
B0	0.275	0.204	0.423	0.000
B1	0.337	0.280	0.423	0.043
B5	0.322	0.263	0.415	0.108

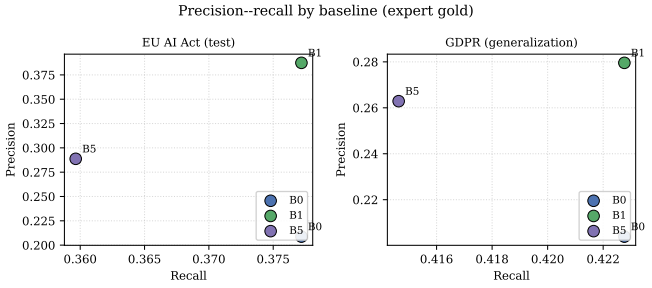

Fig. 2. Precision–recall landscape (B0, B1, B5).

higher hallucination (0.108 vs 0.043 for B1). Figures 5 and 6 (GDPR panels) mirror this ordering; Figure 2 (right) places B1 northeast of B5 and B0 in precision–recall space. Thus the negative answer to our RQ is not an artifact of a single regulation.

C. Trade-offs and failure modes

Figures 2–8 visualize quality–cost and error structure. B1 occupies a favorable position: best F1 at one LLM call per chunk (Figure 3), whereas B5 uses three calls without surpassing B1. Figure 2 places B1 northeast of B5 and B0; Figure 4 shows B5 worse than B1 on both F1 and hallucination despite additional processing. Figures 5 and 6 echo Tables II and III on both corpora. Figure 7 peaks at B1, not at the deepest pipeline; Figure 8 links metrics to failure modes such as invented modality and missed obligations.

a) Summary of tables.: Table II is the primary evidence for the RQ: B1 dominates on F1 while B5 incurs three times the inference calls. Table III confirms that B1's lead is not EU-AI-Act-specific. Together, the tables show that structured decomposition in one call (B1) outperforms role-separated decomposition (B5) on this benchmark, although both beat zero-shot B0.

b) Summary of figures.: Figure 2 shows B1 northeast of comparators in precision–recall space; Figure 3 plots F1 against LLM calls, placing B1 on the quality frontier at one call. Figure 4 highlights that B5 is dominated on both F1 and hallucination relative to B1. Figures 5 and 6 summarize the same rankings on both corpora. Figure 7 makes the non-monotonic effect of pipeline depth visible: adding agents beyond B1 does not increase F1. Figure 8 links metrics to failure modes—B5 exhibits more invented modality (22 vs 8 for B1) and more missed obligations (135 vs 122), suggesting QA removes or distorts valid content.

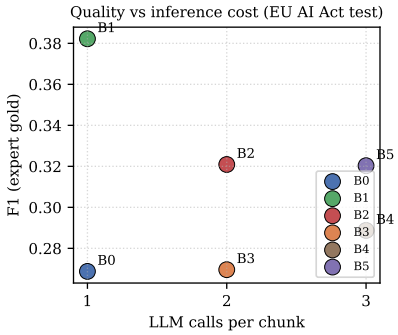

Fig. 3. F1 vs. LLM calls per chunk (B0–B5 ablation).

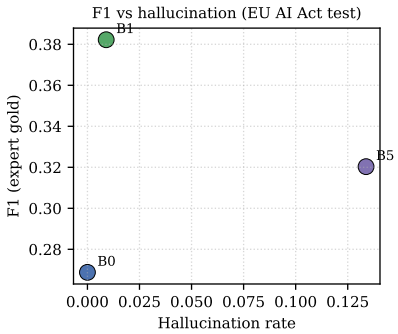

Fig. 4. F1 vs. hallucination rate trade-off (EU AI Act test).

VI. Discussion

For sustainable governance of AI systems, organizations need elicitation workflows that minimize unsupported obligations and inference cost. Matched single-call elicitation (B1) uses the same role text as B5 but avoids error compounding on Llama 3 8B. We answer the primary research question in the negative for matched prompt content: role-separated inference (B5) does not beat the single-call structured prompt (B1) on F1. On EU AI Act test, the QA stage does not reduce hallucination versus the two-agent ablation (B2): B5 hallucination rate exceeds B2, consistent with self-verification limits. Implications: (i) default to B1 for local regulatory RE; (ii) deploy B5 when analysts require constraints/drafts for audit trails; (iii) invest in stronger verifiers or human gold before claiming multi-agent gains. Scores reflect automated expert gold, not legal verdicts.

VII. Conclusion

We evaluated whether a three-agent regulatory RE pipeline (B5) outperforms a matched single-call structured prompt (B1) when both share one local Llama 3 8B model on EU AI Act and GDPR chunks with quote-validated expert gold. B1 achieves the highest F1 on both corpora (0.382 EU test; 0.337 GDPR), ahead of B5 (0.320, 0.322) at one third of the inference cost and with lower hallucination; B5 still beats unstructured B0. Role-separated orchestration therefore does not improve requirement-level fidelity under matched prompts on consumer hardware at this scale.

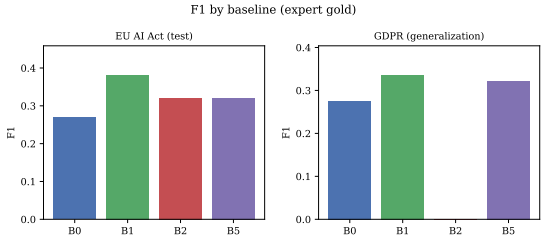

Fig. 5. F1 comparison across baselines (EU AI Act test and GDPR generalization).

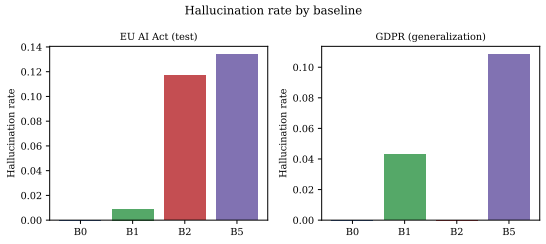

Fig. 6. Hallucination rate comparison across baselines (EU AI Act test and GDPR generalization).

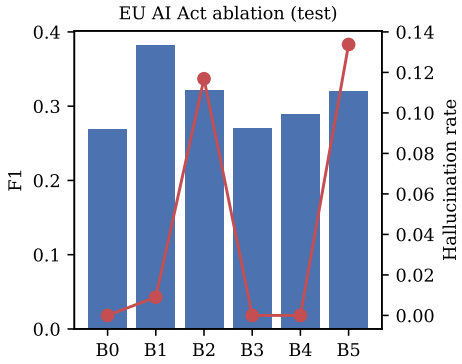

Fig. 7. EU AI Act ablation (B0–B5).

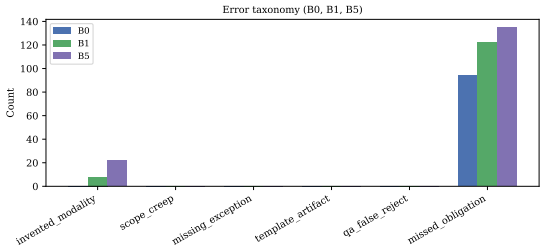

Fig. 8. Error taxonomy counts (B0, B1, B5).

For trustworthy local RE, practitioners should default to structured single-call elicitation; multi-agent workflows are better justified when reviewers need inspectable intermediate outputs than when F1 alone is the goal. Future work should combine stronger verifiers, human-adjudicated gold, and larger models—not additional identical 8B agent roles without added capacity.

References

[1] "Regulation (EU) 2024/1689 (Artificial Intelligence Act)," 2024, oJ L, 12.7.2024.

[2] "Regulation (EU) 2016/679 (General Data Protection Regulation)," 2016.

[3] C. Arora, J. Grundy, and M. Abdelrazek, "Advancing requirements engineering through generative AI: Assessing the role of LLMs," in *Generative AI for Effective Software Development*, A. Nguyen-Duc, P. Abrahamsson, and F. Khomh, Eds. Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 129–148.

[4] "CrewAI: Multi-agent orchestration framework," https://www.crewai.com, 2024.

[5] M. A. Zadenoori, J. Dąbrowski, W. Alhoshan, L. Zhao, and A. Ferrari, "Large language models (LLMs) for requirements engineering (RE): A systematic literature review," 2025. [Online]. Available: https://arxiv.org/abs/2509.11446

[6] "Ollama: Local LLM runtime," https://ollama.com, 2024.

[7] Anonymous, "Anonymous-by-construction: An LLM-driven framework for privacy-preserving text," 2026. [Online]. Available: https://arxiv.org/abs/2603.17217

[8] C. Zhang *et al.*, "Rethinking the value of multi-agent workflow: A strong single agent baseline," 2026. [Online]. Available: https://arxiv.org/abs/2601.12307

[9] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4310–4330. [Online]. Available: https://aclanthology.org/2022.acl-long.297

[10] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang *et al.*, "AgentBench: Evaluating LLMs as agents," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024, arXiv:2308.03688. [Online]. Available: https://openreview.net/forum?id=SIthA1gQ52

[11] R. Etezadi, S. Abualhaija, C. Arora, and L. Briand, "Classifier or prompt: A case study on legal requirements traceability," 2025. [Online]. Available: https://arxiv.org/abs/2502.04916

[12] S. Abualhaija, M. Ceci, N. Sannier, D. Bianculli, S. Lannier, M. Siclari, O. Voordeckers, and S. Tosza, "LLM-assisted extraction of regulatory requirements: A case study on the GDPR," in *Proceedings of the 33rd IEEE International Requirements Engineering Conference (RE)*. IEEE, 2025.

[13] Y. Liu, B. Chong, Y. Lin, C. Zhang, H. Zheng, Z. Zhang, J. Liang, R. Ran, Q. Li, and K. Xu, "Reducing detail hallucinations in long-context regulatory understanding via targeted preference optimization," 2026. [Online]. Available: https://arxiv.org/abs/2604.23113

[14] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou, "Large language models cannot self-correct reasoning yet," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://openreview.net/forum?id=IkmD3fKBPQ

[15] J. O. Agyemang, J. J. Kponyo, E. Amponsah, G. M. A. Boakye, and K. O.-B. O. Agyekum, "LLM-redactor: An empirical evaluation of eight techniques for privacy-preserving LLM requests," 2026. [Online]. Available: https://arxiv.org/abs/2604.12064