

Fiscalização Inteligente: Um Assistente Virtual Jurídico-Administrativo Baseado em LLM para Suporte à Gestão de Contratos – Um Relato Prático

Daniel B. S. Sousa¹, Murilo M. Q. C. Santos¹, Ludmila B. Silva¹, Cassiano H. Albuquerque², Emily B. S. Pinto³

¹Coordenação de Sistemas de Informação – Instituto Federal de Alagoas (IFAL)
Campus Arapiraca – Rodovia AL-110, 359, Deputado Nezinho – 57317-291 –
Arapiraca – AL – Brasil

²Instituto Federal de Alagoas (IFAL) - Campus Maragogi – Rodovia AL-101, 139, –
57955-000 – Maragogi – AL – Brasil

³Instituto Federal de Alagoas (IFAL) - Reitoria – Rua Odilon Vasconcelos, 103, –
57035-660 – Maceió – AL – Brasil

{dbss1,mmqcs1,lbs11}@aluno.ifal.edu.br, {cassiano.henrique,
emily.santos}@ifal.edu.br

Abstract. *Public contract management requires swift interpretation of complex regulations. In a federal multi-campus institution, decentralization hinders support for inspectors across geographically dispersed units, increasing administrative risks. This paper presents GeFi (Gestão e Fiscalização Inteligente), a virtual assistant based on Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). The objective is to provide support in interpreting institutional documents, aiming to enhance agility and legal certainty in contract management. Partial results demonstrate the potential of generative AI for knowledge centralization and digital transformation in the public sector.*

Resumo. *A gestão de contratos na administração pública demanda interpretação célere de normas complexas. Em uma instituição federal multicampi, a descentralização dificulta o suporte a fiscais distribuídos em unidades geograficamente dispersas, ampliando riscos administrativos. Este trabalho apresenta o GeFi (Gestão e Fiscalização Inteligente), assistente virtual fundamentado em Large Language Models (LLMs) e arquitetura Retrieval-Augmented Generation (RAG). O objetivo é prover suporte na interpretação de documentos institucionais, visando ampliar a agilidade e a segurança jurídica na gestão contratual. Resultados parciais demonstram o potencial da IA generativa na centralização do conhecimento e na transformação digital do setor público.*

1. Introdução

A gestão e a fiscalização de contratos constituem atividades críticas na administração pública, pois envolvem execução orçamentária, conformidade normativa e responsabilização de agentes públicos [Matos and Reis 2023; Silva et al. 2024]. Em organizações multicampi, a equipe central precisa atender unidades geograficamente dispersas, responder a dúvidas recorrentes e interpretar documentos heterogêneos, leis, normativos internos, pareceres, manuais e contratos [Barbosa et al. 2024; STJ 2023].

No contexto deste relato, a unidade responsável atende cerca de cem fiscais distribuídos em dezessete campi e responde por aproximadamente setenta por cento do orçamento de custeio institucional. A recuperação de informações é lenta e depende de conhecimento especializado, o que gera retrabalho, inconsistências e aumento de riscos administrativos [STJ 2023; Barbosa et al. 2024].

Diante desse problema, foi concebido o GeFi, assistente virtual jurídico-administrativo baseado em LLMs e arquitetura RAG [Lewis et al. 2020; Gao et al. 2023; Zhao et al. 2024]. O desenvolvimento seguiu a Design Science Research Methodology (DSRM) [Peppers et al. 2007], percorrendo as etapas de (1) identificação do problema, (2) definição de objetivos, (3) design e desenvolvimento do artefato e (4) demonstração. A fase de avaliação empírica estruturada está em planejamento para ciclos futuros. A proposta oferece respostas contextualizadas, ancoradas em corpus normativo atualizado, ampliando a controlabilidade e reduzindo riscos de alucinação.

A experiência dialoga com iniciativas recentes no LASDiGov, como o desenvolvimento de chatbots baseados em IA generativa para apoio à informação acadêmica em instituições públicas, reforçando o potencial de assistentes especializados para demandas institucionais [Layane et al. 2025]. Essa tendência é apoiada pelo desenvolvimento do assistente Sabiá [Rodrigues, G. B. et al. 2025], que utiliza a arquitetura RAG para mitigar a fragmentação de dados acadêmicos e simplificar o acesso a normativas e resoluções no ensino superior. Os dois cenários mostram como o uso de interfaces de consulta inteligente pode transformar a governança da informação no setor público, convertendo documentos complexos e descentralizados em respostas claras e relevantes para os usuários.

2. Situação e Contexto

A demanda surgiu da rotina da coordenação de gestão e fiscalização contratual. As dúvidas dos fiscais envolviam prazos, procedimentos, competências, limites de atuação, tratamento de ocorrências e interpretação de orientações internas. Embora muitas respostas já estivessem previstas em documentos institucionais, sua recuperação dependia de servidores experientes e era lenta.

Esse cenário evidenciou quatro problemas centrais: (1) dispersão informacional em diferentes arquivos; (2) assimetria de conhecimento entre equipe central e os campi; (3) sobrecarga operacional com dúvidas repetitivas; e (4) risco de inconsistência interpretativa em respostas elaboradas manualmente.

A solução deveria atender a requisitos mínimos: fundamentação em fontes rastreáveis, atualização da base sem retreinamento, preservação de logs de consulta, controle de acesso e manutenção do humano no circuito decisório [Gao et al. 2023; Zhao et al. 2024]. Por se tratar de aplicação no setor público, foram considerados aspectos de governança da informação, como minimização do envio de dados sensíveis e definição de sua natureza assistiva, voltada ao apoio à decisão, não à automação de decisões administrativas [Wu et al. 2025; Rodrigues, G. J. A. 2025].

3. Ações Realizadas

A primeira ação foi o levantamento e validação de requisitos por meio de reuniões com fiscais, gestores e equipe de desenvolvimento, definindo requisitos funcionais e não funcionais com ênfase em consultas em linguagem natural, rastreabilidade, segurança e usabilidade.

A segunda ação concentrou-se na organização da base de conhecimento. A equipe reuniu normas, manuais, orientações internas e outros documentos relevantes, submetendo-os a processo de seleção, padronização e definição de metadados. A qualidade da recuperação depende diretamente da qualidade do corpus indexado [Gao et al. 2023; Zhao et al. 2024].

A terceira ação foi a definição da arquitetura modular, com separação entre frontend (Next.js 14 e Tailwind CSS), backend (NestJS e Prisma), persistência (PostgreSQL + pgvector) e camada de recuperação semântica. O pipeline de RAG opera em três fases: (i) ingestão, documentos PDF são segmentados em partes menores (*chunks*); (ii) indexação, onde cada *chunk* é convertido em representação vetorial utilizando modelos de *embedding* do Google Gemini e armazenado no banco pgvector com metadados de rastreabilidade (documento, data); e (iii) consulta, em que a pergunta do usuário é codificada no mesmo espaço vetorial, permitindo que os trechos mais relevantes por similaridade sejam recuperados e enviados como contexto ao modelo de linguagem (Google Gemini). A Figura 1 sintetiza essa arquitetura (blocos 1–11 correspondem, respectivamente, à pergunta do usuário, requisição à API, geração do embedding da query, busca semântica, consulta ao banco vetorial, recuperação de trechos, montagem do contexto, envio à LLM, resposta formatada e retorno ao frontend).

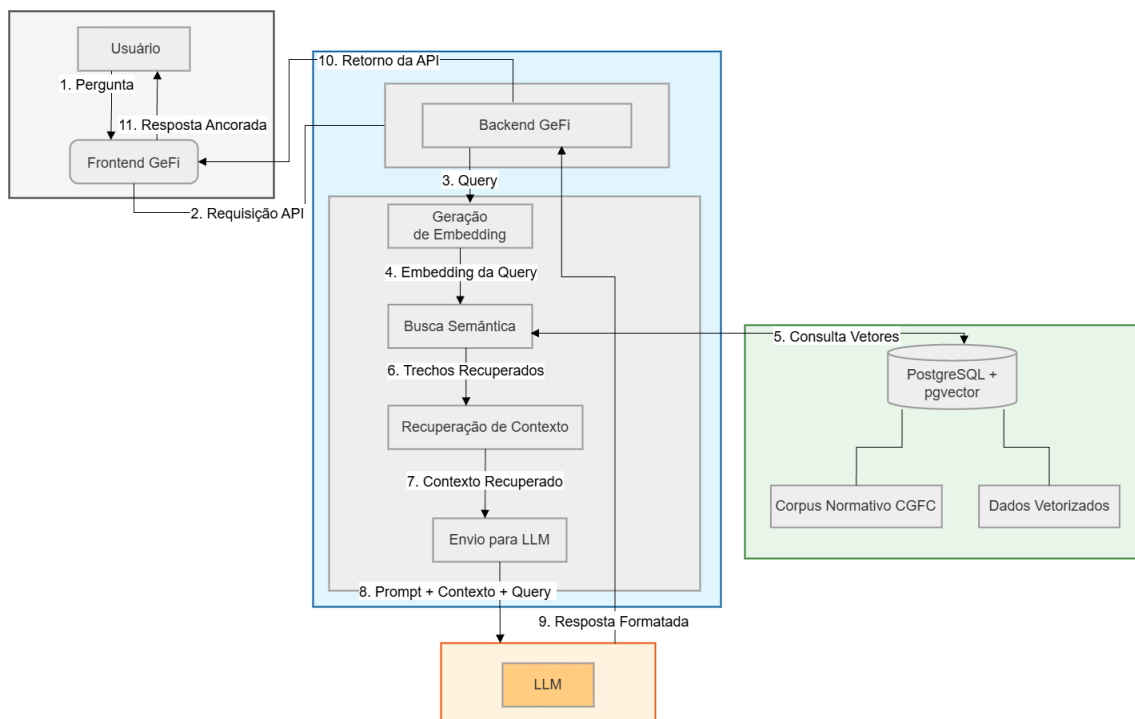


Figura 1. Arquitetura técnica do agente de IA com RAG

A quarta ação foi a prototipação de alta fidelidade. Antes da implementação completa, foram desenhadas telas para login, interação em formato de chat e gestão da biblioteca documental. A prototipação permitiu alinhar expectativas com os futuros usuários, ajustar nomenclaturas institucionais e reduzir ambiguidades de navegação. Essa etapa também ajudou a delimitar melhor o papel da ferramenta: usuários não buscavam um chatbot genérico, mas um instrumento objetivo, com respostas claras, rastreáveis e visualmente compatíveis com a linguagem administrativa.

A quinta ação foi o início da implementação material do sistema. Foram estruturados repositórios separados para documentação, design, frontend e backend, com uso de Git e GitHub para controle de versão. A codificação avançou em paralelo à revisão de requisitos e dos protótipos, o que exigiu revisitações frequentes a decisões anteriores. Também foram elaborados diagramas de casos de uso, diagrama de classes e modelo lógico de banco de dados, não como artefatos meramente formais, mas como instrumentos de comunicação entre as frentes de desenvolvimento e de negócio.

Por fim, a equipe incorporou decisões de governança ao desenho da solução. Foi definido que a ferramenta deveria apresentar referências explícitas às fontes recuperadas, preservar logs mínimos de uso para fins de auditoria e diferenciar perfis de acesso conforme o papel institucional do usuário. Também se buscou reduzir o tratamento desnecessário de dados pessoais, priorizando no corpus inicial documentos normativos, orientações internas e materiais administrativos de circulação institucional. Essas escolhas não eliminam riscos, mas estabelecem uma base mais prudente para uso assistivo de IA no setor público [Wu et al. 2025; Rodrigues, G. J. A. 2025].

4. Resultados

Os resultados alcançados são parciais, mas já evidenciam a materialização inicial do artefato. O primeiro resultado foi a transformação de uma demanda difusa, antes baseada em dúvidas esparsas e consultas manuais, em um problema organizacional delimitado e traduzido em requisitos de sistema, com escopo, usuários e objetivos definidos.

O segundo resultado foi a consolidação de uma base documental inicial voltada à gestão e à fiscalização contratual. Ainda que o processo de curadoria permaneça em evolução, a simples centralização do corpus já representou um avanço organizacional. Documentos que antes estavam dispersos em diferentes repositórios passaram a compor uma biblioteca digital unificada, com critérios mínimos de seleção, armazenamento e administração desse acervo.

O terceiro resultado foi a definição de uma arquitetura técnica coerente com as restrições do problema. A separação entre interface de usuário (frontend), a lógica de negócio (backend), a persistência em banco de dados e a camada de recuperação semântica permitiu isolar responsabilidades e facilitar futuras substituições tecnológicas. Essa decisão mostrou utilidade concreta durante ajustes feitos na camada de persistência

e na organização do backend, que puderam ocorrer sem reescrita integral das demais partes do sistema. Para um ambiente institucional sujeito a mudanças de infraestrutura e de equipe, essa modularidade tende a favorecer a manutenção e a evolução da solução.

O quarto resultado foi a construção de protótipos de alta fidelidade capazes de representar o fluxo principal da aplicação. A navegação prevista contempla autenticação, interação em linguagem natural e administração da base de documentos. A prototipação contribuiu para avaliar preliminarmente a adequação da interface à rotina dos usuários e reduziu o risco de se desenvolver uma solução tecnicamente sofisticada, porém desalinhada das expectativas operacionais.

O quinto resultado foi a implementação inicial do sistema, com avanço simultâneo em frontend, backend e modelagem de dados. Embora a solução ainda não esteja apresentada como produto final implantado em produção, já existe uma base funcional capaz de orientar os próximos ciclos de refinamento. Esse estágio é particularmente importante porque desloca o projeto do campo da intenção para o campo da execução verificável. A Figura 2 apresenta a interface principal de interação já prototipada e implementada em versão funcional inicial, ilustrando o fluxo de consulta em linguagem natural e a apresentação de respostas assistivas ao usuário.

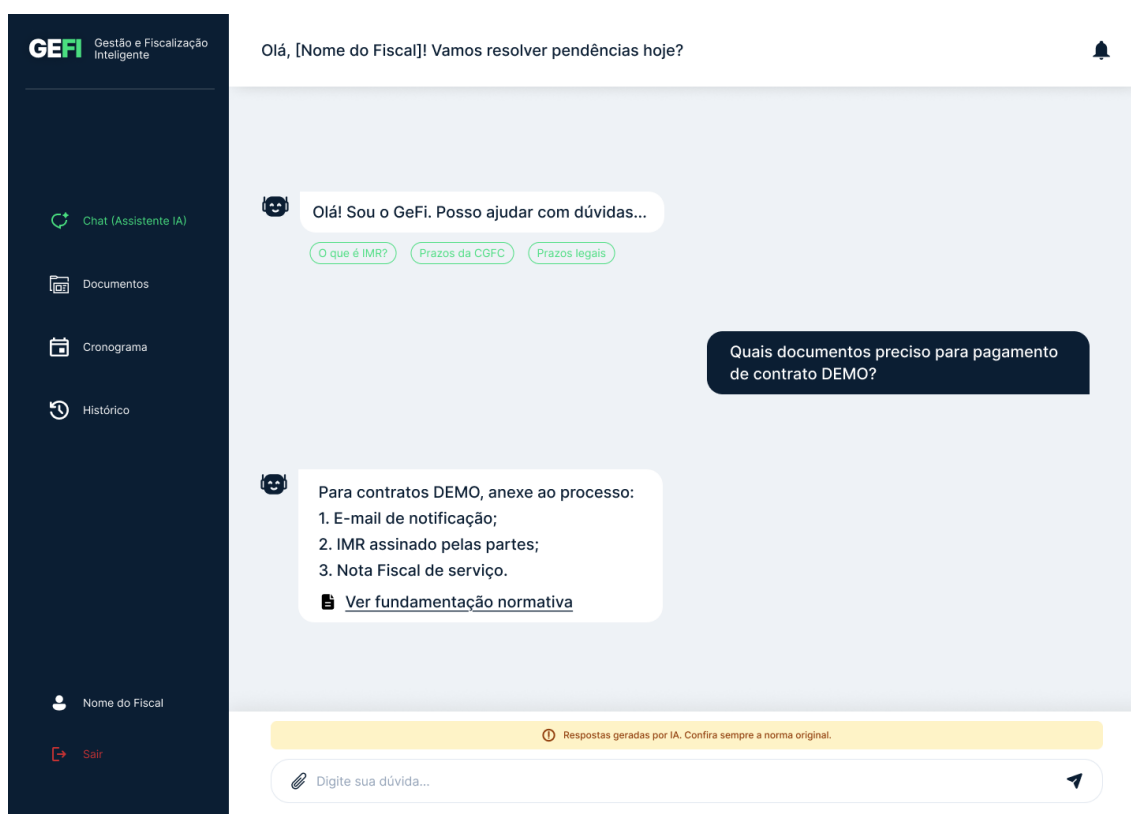


Figura 2: Interface principal do GeFi com interação em linguagem natural e resposta assistiva ao usuário.

Além dos artefatos técnicos, a discussão sobre o GeFi estimulou a explicação de rotinas, a revisão do acervo documental e a reflexão sobre responsabilidades e limites de

uso da IA na gestão contratual. Mesmo antes da implantação plena, o projeto já atua como indutor de padronização e aprendizagem institucional.

4.1. Limitações do estudo

A avaliação qualitativa realizada é exploratória e não permite generalizações. Três limitações principais merecem destaque. Primeira: a precisão do sistema depende criticamente da qualidade do corpus, diante de normativos conflitantes, documentos desatualizados ou perguntas ambíguas, o modelo pode apresentar alucinações, gerando respostas linguisticamente coerentes, porém juridicamente imprecisas [Gao et al. 2023; Zhao et al. 2024]. Segunda: o risco de confiança excessiva por parte dos usuários é real, a fluência da resposta pode induzir o fiscal a dispensar a verificação na fonte original, o que o design da interface procura mitigar, sem eliminar completamente.

5. Lições Aprendidas

A primeira lição aprendida é que projetos com LLMs no setor público enfrentam gargalos que vão muito além das limitações tecnológicas, esbarrando essencialmente na governança da informação. A curadoria dos dados revelou-se o maior desafio do projeto: definir versões corretas de documentos, padronizar metadados e estabelecer critérios rígidos de atualização são aspectos institucionais que condicionam diretamente a confiabilidade percebida da solução [Gao et al. 2023; Zhao et al. 2024]

A segunda lição é que a adoção da arquitetura RAG ultrapassa a justificativa técnica; ela é uma exigência de accountability. Em contexto de gestão contratual, onde respostas genéricas são administrativamente inaceitáveis, a capacidade de rastrear e ancorar a resposta gerada às suas fontes documentais é o principal mecanismo para viabilizar o uso seguro de IA generativa em ambientes regulados [Lewis et al. 2020; Gao et al. 2023].

A terceira lição é que a prototipação não deve ser tratada como etapa cosmética. O desenho das interfaces funcionou como mecanismo de mediação entre linguagem técnica e linguagem administrativa, muitos ajustes importantes surgiram na visualização concreta de telas, fluxos e nomenclaturas, não durante a codificação.

A quarta lição é que a promessa de automação precisa ser contida. Em vez de apresentar o GeFi como sistema autônomo de decisão, a equipe obteve melhor alinhamento institucional ao posicioná-lo como assistente de consulta e apoio interpretativo. Essa escolha reduz riscos, preserva a responsabilidade humana e favorece a aceitação do projeto em um ambiente em que segurança jurídica e prestação de contas são centrais [Rodrigues, G. J. A. 2025].

A quinta lição é que privacidade e segurança não podem ser anexos de última hora. Mesmo com corpus composto majoritariamente por normas e documentos administrativos, é necessário prever controle de acesso, registro de uso, administração da biblioteca e critérios para inclusão ou exclusão de materiais desde a concepção [Wu et al. 2025].

Como próximos passos, a equipe planeja: (i) ampliar o corpus e realizar avaliação formal com grupo de controle para medir redução de erros processuais em relação à busca manual, avaliando utilidade percebida, clareza das respostas, aderência às fontes e tempo de consulta; e (ii) explorar a integração a sistemas de processos eletrônicos para geração automática de check-lists e minutas de encaminhamento ao iniciar uma contratação [Peffer et al. 2007].

6. Recomendações para Replicação

A experiência permite indicar cuidados mínimos para organizações públicas que pretendam adotar assistentes baseados em LLMs e RAG em contextos regulados. Recomenda-se iniciar pelo mapeamento do problema organizacional, pela definição do corpus documental e pela delimitação explícita do papel da ferramenta. Em contextos jurídico-administrativos, soluções mal posicionadas tendem a ampliar riscos organizacionais e interpretativos, especialmente quando não há clareza sobre fontes, critérios de atualização e limites de uso.

Recomenda-se ainda que a implantação ocorra de forma gradual, com acompanhamento das dúvidas mais recorrentes, revisão periódica do acervo e um canal explícito para correção de respostas problemáticas. A credibilidade da solução não depende apenas do desempenho técnico, mas também da capacidade institucional de manter a base atualizada, corrigir erros, preservar a supervisão humana e comunicar com transparência os limites de uso da ferramenta.

7. Referências

- Barbosa, R. F.; Lima, T. C.; Brito, G. K. L.; Marques, J. F. S.; Lucas, O. G.; Silva Neto, A. F.; Silva, A. N. (2024) “Fiscalização de contratos nas instituições federais de ensino superior: desafios e perspectivas na gestão contratual”. *Revista Foco*, v. 17, n. 12, e7100, p. 1–19. DOI: 10.54751/revistafoco.v17n12-033.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. (2023) “Retrieval-Augmented Generation for Large Language Models: A Survey”. *arXiv preprint arXiv:2312.10997*.
- Layane, L.; Machado, G.; Morandi, G.; Silva, N.; Junior, W.; Alves, E. (2025) “Chatbot customizável baseado em IA generativa para informação de conteúdos acadêmicos de instituições de ensino superior públicas”. In: *Anais do XII Latin American Symposium on Digital Government (LASDiGov 2025)*, p. 169–180. Porto Alegre: Sociedade Brasileira de Computação. doi:10.5753/lasdigov.2025.9016.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; Kiela, D. (2020) “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. *arXiv preprint arXiv:2005.11401*.
- Matos, F. R. P.; Reis, D. L. (2023) “Gestão de contratos e sua aplicação como instrumento de redução de custo: planejamento e fiscalização dos recursos públicos”.

Revista Científica Semana Acadêmica, v. 11, n. 232. DOI: 10.35265/2236-6717-232-12536.

Peppers, K.; Tuunanen, T.; Rothenberger, M. A.; Chatterjee, S. (2007) “A Design Science Research Methodology for Information Systems Research”. *Journal of Management Information Systems*, v. 24, n. 3, p. 45–77. DOI: 10.2753/MIS0742-1222240302.

Rodrigues, G. B., Beal, F., Marcon, M., Souza, A. C. C., Ortoncelli, A. R., Souza, F. C. M., and Silva, R. A. (2025). "Sabiá: Um Chatbot de Inteligência Artificial Generativa para Suporte no Dia a Dia do Ensino Superior". In: Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025). p. 1401-1414.

Rodrigues, G. J. A. (2025) “Aplicações da inteligência artificial no setor público: impactos na produtividade e nos processos administrativos”. Trabalho de Conclusão de Curso, Instituto Federal da Paraíba.

Silva, E. C.; Moreira, A. A. A. P.; Nascimento, L. R.; Serafim, G. L.; Santos, H. S.; Mendes, M. C. (2024) “Uma revisão sistemática do uso de tecnologias digitais na modernização e agilização dos processos de licitação pública”. *Revista Foco*, v. 17, n. 2, e4491, p. 1–20. DOI: 10.54751/revistafoco.v17n2-106.

STJ (2023) *Manual de Gestão e Fiscalização de Contratos*. Brasília: Superior Tribunal de Justiça.

Wu, Z.; Guo, J.; Hou, J.; He, B.; Fan, L.; Yang, Q. (2025) “Model-based Large Language Model Customization as Service”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, p. 4904–4921. DOI: 10.18653/v1/2025.emnlp-main.248.

Zhao, S.; Yang, Y.; Wang, Z.; He, Z.; Qiu, L. K.; Qiu, L. (2024) “Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely”. *arXiv preprint arXiv:2409.14924*.