



K-CAT: AN AI-POWERED FRAMEWORK FOR DOMAIN-AWARE ENTERPRISE MACHINE TRANSLATION

OFF-THE-SHELF LARGE LANGUAGE MODELS
TRANSLATE WORDS. K-CAT TRANSLATES INTENT.

Abstract

Translation in certain fields such as law and finance requires more than just linguistic competence. A lack of context in interpreting domain-specific terminology can result in mistranslation, altering legal or regulatory meaning. This, in turn, can lead to compliance risks or legal complications. In this context however, translation is not just a language consideration; it becomes a risk management challenge.

This paper introduces Knowledge and Context Augmentation for Domain-aware Translation (K-CAT), an enterprise framework for the legal domain that embeds legal terminology, domain-specific definitions, and bilingual precedents into translation pipelines. By integrating structured domain knowledge into the translation process, K-CAT ensures accurate, explainable, and audit-ready translations.

The paper discusses how K-CAT is used as a benchmark for generic large language model (LLM) translations and other leading commercial translation software. It aims to demonstrate how integrating domain-specific knowledge, and not model scale alone, can effectively bridge domain-specific accuracy gaps.

Introduction

Large enterprises operate across geographies and multiple jurisdictions, where official documents including contracts, regulatory filings, and cross-border agreements must be translated accurately into different languages. These documents often require high-quality, domain-aligned translations.

Recently, generative artificial intelligence (GenAI) has become the popular approach for document translations. However, multilingual large language models (LLMs), often generic and independent of tasks or domains, do not readily produce domain-aware translations. They are trained for linguistic fluency, produce semantically correct translations, but are not domain accurate. Commercial machine translation (MT) systems also have similar limitations.

To address this gap, Infosys has developed Knowledge and Context Augmentation for Domain-aware Translation (K-CAT), a framework that incorporates domain grounding in the translation process. K-CAT incorporates the enterprise contract ontology (ECO), a continuously validated, human-curated layer of domain definitions, usage examples, and bilingual precedents. This layer is embedded into the LLM's context before translation to enhance model accuracy.

Experimental results on a 10,000-sentence English–German legal corpus spanning 20 legal categories demonstrate that K-CAT improves translation accuracy by about 73% over a baseline generic LLM, using the Bilingual Evaluation Understudy (BLEU) translation metric. It also consistently outperforms commercial translation tools.

This paper highlights K-CAT augmentation steps and strategies, as well as improvements recorded at each stage.



Domain-aware Translation: Interpretation Risks Vs. Language Fluency

Official documents often contain specialized terminology, jurisdiction-specific nuances, and regulatory language that differ from everyday usage. While generic LLMs understand language fluently, they frequently fail to capture the intent and domain-specific meaning of specialized terminologies.

This is because LLMs are optimized for broad linguistic coverage rather than domain-specific accuracy. Consequently, translating contracts, regulatory filings, or vendor agreements without domain context can lead to ambiguity, compliance risks, increased human review effort, and, where necessary, additional manual rework to mitigate potential legal consequences.

Traditional statistical and neural MT systems rely on large volumes of general-purpose parallel data. They cannot be scaled to specialized domains, where high-quality, legally validated parallel corpus is limited. LLMs inherit the same limitations as they are primarily trained on general-purpose datasets. They possess broad language fluency but have shallow domain expertise in areas such as enterprise contracting.

Table 1 cites examples that illustrate the importance of domain awareness in legal translations. Many common terms have specialized legal meanings that differ significantly from their everyday usage. A translation engine that relies on the common interpretation may produce a fluent but incorrect translation, altering the intended meaning of the contractual clauses.

Table 1: Interpretation differences in legal and everyday terminology

Term	General Meaning	Legal Meaning
Consideration	Thoughtfulness	Contractual value exchanged
Party	Celebration	Contracting entity
Execute	Run	Legally sign
Material	Substance	Significant contractual impact

These examples highlight the gap between language fluency and domain-specific accuracy. The core premise of K-CAT is reducing the gap between language fluency and domain accuracy, while producing measurable improvements.

With manually validated definitions, K-CAT achieved a 73% BLEU score improvement over a baseline generic LLM, ensuring a scalable and enterprise-ready translation pipeline.

K-CAT Methodology

K-CAT is an enterprise framework that systematically integrates domain knowledge and context into the translation process rather than relying on LLM translation capabilities.

We evaluated nine K-CAT methods across four augmentation approaches to determine the most effective type of embedded domain context, as summarized in Table 2. The evaluation was conducted on an English to German legal translation corpus to substantiate the merit of this process.

Table 2: K-CAT augmentation families and corresponding methods

Augmentation Family	Method	Translation Strategy
Instruction augmentation (IA)	Zero-shot prompting (ZSP)	Baseline instruction-only translation without domain context
	Domain-aware prompting (DAP)	Explicit framing of the text as a legal or contract domain task
	Reasoning-driven prompting (RDP)	Reasoning of term meanings before translation
Demonstration augmentation (DA)	Semantic example retrieval (SER)	Retrieves semantically similar translated examples as few-shot context
	Category-guided example retrieval (CGER)	Retrieves examples from the same legal category as the input
Knowledge augmentation (KA)	Terminology-guided translation (TGT)	Injects validated glossary terms relevant to the input sentence
	Definition-enhanced terminology guidance (DETG)	Embeds complete legal term definitions and usage
Hybrid augmentation (HA)	Category-aware hybrid guidance (CAHG)	Integrates category-guided examples with terminology and definitions
	Term-centric hybrid guidance (TCHG)	Combines term-level definitions, usage guidance, and bilingual precedents



K-CAT Translation Workflow

At the core of K-CAT is the ECO, a dynamic legal knowledge layer comprising validated definitions and authentic usage examples of domain-specific terminology. The ECO is built

and updated through a human-in-the-loop (HITL) pipeline rather than relying only on model inference. This ensures K-CAT's output is explainable and audit-ready.

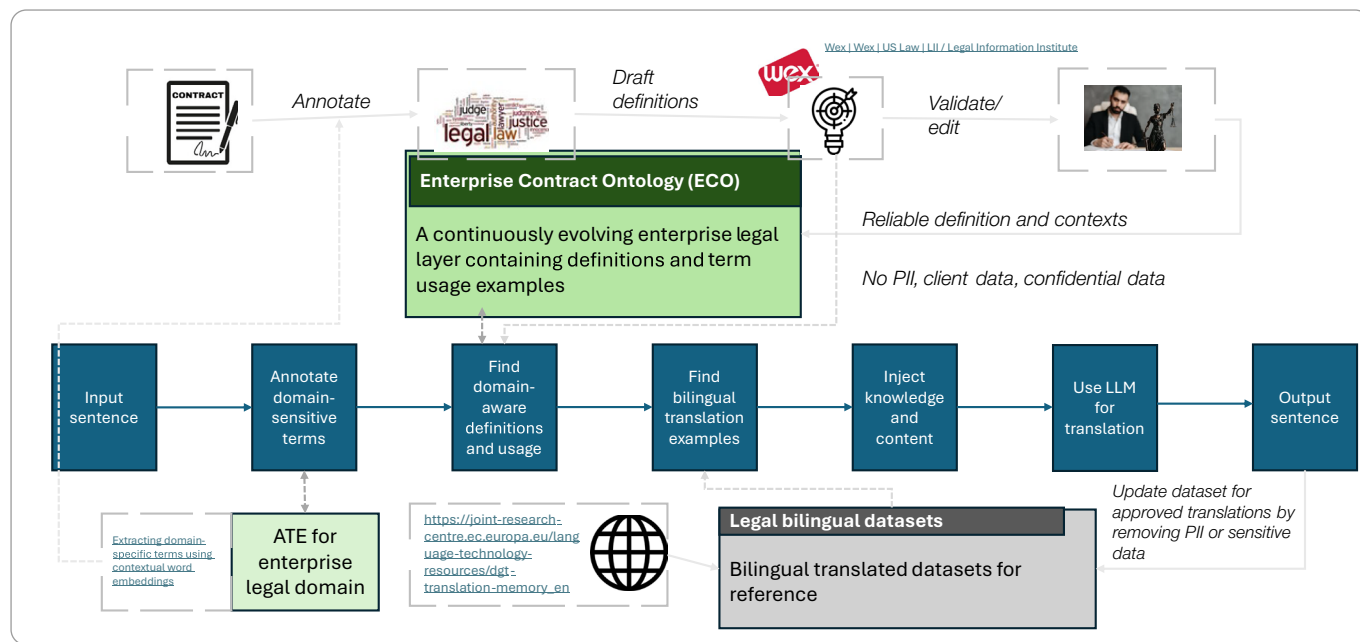


Fig 1: K-CAT workflow: Knowledge-build and runtime loops

As depicted in Figure 1, K-CAT involves two interconnected workflows: the knowledge-build flow and runtime translation flow.





Knowledge-build flow

Input contracts are annotated for domain-specific terminology using automatic term extraction (ATE). This technique is used for identifying relevant terms and phrases from text, tailored to the enterprise legal domain for this study. Legal terminology often includes multi-word expressions, ranging from single words to phrases with five or six words. A robust ATE¹ algorithm is therefore essential to interpret the terminology accurately. Extracted terms are then validated against authentic legal reference sources, such as Wex, the legal dictionary maintained by Cornell Law School's Legal Information Institute, to generate draft definitions.² Experts review and validate these definitions before the terms are incorporated into the ECO.



Runtime flow

Each input sentence is first annotated for domain-sensitive terminology. It is then matched against the ECO for reliable definitions and usage examples. The terms are further compared with legal bilingual datasets, such as the EU's Directorate-General for Translation-Translation Memory (DGT-TM) for relevant bilingual precedents³. This domain knowledge and context is embedded into LLM prompts before translation, ensuring accurate interpretations of context-sensitive legal terms.

A key design principle of K-CAT is that no personally identifiable information (PII), client data, or confidential information is stored in the ECO or the bilingual dataset. Approved translations are incorporated into the datasets only after sensitive data has been removed. This enables continuous improvement of the knowledge base from real-world use cases without compromising data privacy and governance requirements.

Experimental Setup and Evaluation

K-CAT was evaluated on a 10,000-sentence English–German legal corpus spanning 20 detected categories. The evaluation leveraged a knowledge base containing 5,553 validated

legal glossary terms across four translation approaches implemented for comparison.

Each of the nine methods was scored using standard, widely-used MT metrics: BLEU and Metric for Evaluation of Translation with Explicit Ordering (METEOR). As illustrated in Figure 2, higher scores indicate closer alignment with reference translations^{4,5}.

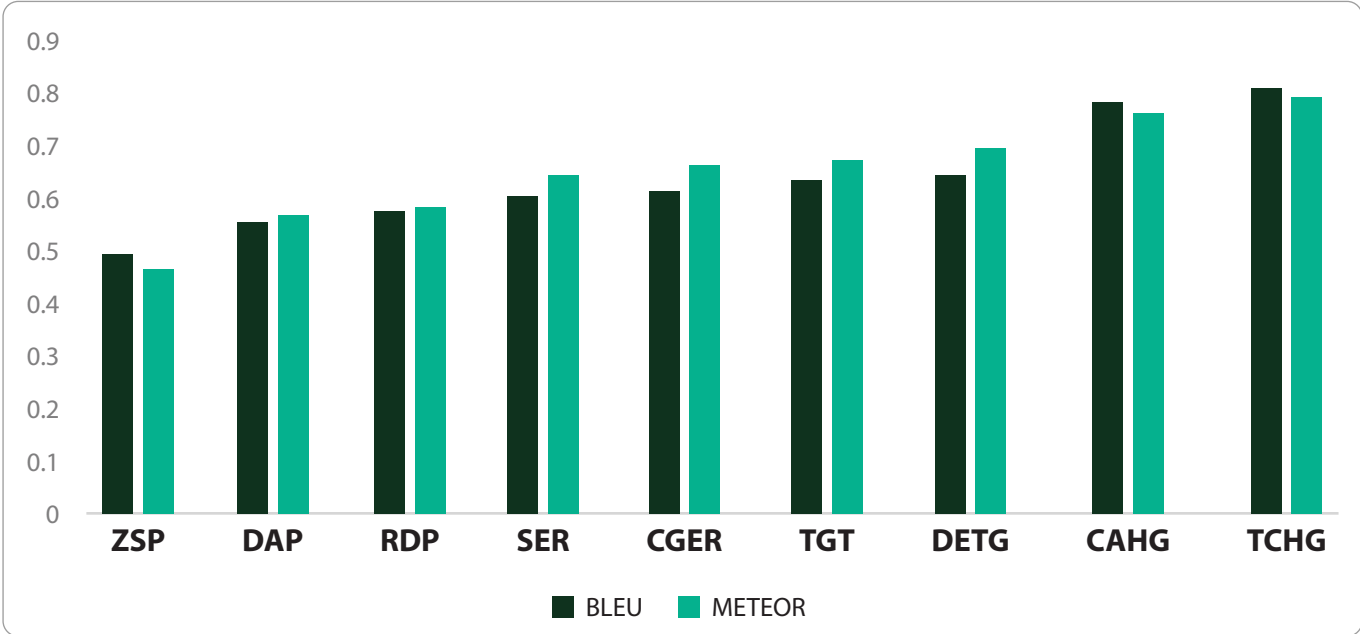


Fig 2: Comparison of BLEU and METEOR scores across K-CAT methods

A common pattern with both metrics is that instruction-only augmentation (ZSP, DAP, RDP) results in smaller improvements. This is because model translations are done without additional domain knowledge. Both demonstration and knowledge augmentation methods produce meaningful outputs. However, hybrid approaches, which combine category-specific examples with term-level definitions and bilingual precedents, consistently deliver an optimized performance.

The TCHG method produces a BLEU score of 0.82 against a generic zero-shot LLM baseline of about 0.47, an improvement of approximately 73%. This was achieved through manually validated definitions in a scalable

pipeline, applied across domains beyond legal contracting.

In addition, we also benchmarked K-CAT against widely used commercial translation tools on the same English–German legal test set. The systems were evaluated using their default, out-of-the-box configuration without domain-specific customization, in line with the current enterprise deployments. Even under these reasonably competitive conditions, K-CAT’s hybrid guidance methodology outperformed commercial translation systems in accurately translating domain-aware terminology and clause-level phrasing, both of which are prone to legal or contractual drifts when translated using off-the-shelf tools.

While commercial translation tools remain a reasonable choice for general correspondence, our results suggest that they should not be a default choice for contracts, regulatory filings, or other legally binding documents. Relying on them highlights accuracy gaps, leading to compliance and rework risks.

K-CAT: Enterprise Applications and Benefits

K-CAT streamlines contract reviews and cross-geographical translations, while minimizing legal errors. By grounding translations in validated domain-specific definitions rather than relying on generic language patterns, K-CAT improves translation consistency and reduces hallucinations. Additionally, it makes translation decisions explainable and traceable, essential for large volumes of cross-language contracts.

K-CAT's knowledge base is enterprise-specific, adaptable, and built entirely on human-validated, non-client-specific terminology. This makes the framework scalable across domains, including global contracting, regulatory and compliance documentation, cross-border operations, vendor management, and cross-jurisdiction mergers and acquisitions (M&A).

Conclusion

Domain-aware translation cannot rely on linguistic fluency alone as it may lead to interpretation risks of a subset of legal terms that generic language models are not designed to address.

K-CAT tackles this challenge by reframing translation as a knowledge-augmentation process, grounding outputs in a continuously validated ECO and legal bilingual precedents. K-CAT achieved 73% improvement over a baseline generic LLM translation, while outperforming commercial translation tools evaluated on the same corpus.

The need for accurate contracting and regulatory documentation continues to grow rapidly across geographies. To safeguard themselves from compliance or legal issues, enterprises should view domain-aware translation as a governed, explainable capability, rather than a generic language utility.



About the Authors



Dr. Prawaal

Principal Data Scientist, Infosys

Dr. Prawaal is a Principal Data Scientist at Infosys with over 25 years of experience in the IT industry. He specializes in AI, natural language processing (NLP), and GenAI, and has published several research papers at leading conferences. He advises enterprises on AI strategies that enhance business value. As an active writer and speaker on AI governance and risks, Dr. Prawaal also contributes to AI for Social Good (AI4SG) initiatives by developing AI models for tribal languages.



Leemamol Mathewkutty

Lead Analyst, Analytics Technology, Infosys

Leemamol Mathewkutty is a Lead Analyst, Analytics Technology, at Infosys. With expertise in AI, data analytics, and enterprise technology solutions, she drives innovation, decision making, process optimization, and data-driven business outcomes across enterprises.



Hitanshu Dadhich

Digital Specialist Engineer, Infosys

Hitanshu Dadhich is a Digital Specialist Engineer at Infosys, specializing in data engineering, AI, machine learning, and agentic AI. With a focus on delivering business value, Hitanshu's work spans designing data-driven solutions, building scalable data pipelines, and developing intelligent AI-powered applications.

References

1. arXiv | [Extracting Domain-Specific Terms Using Contextual Word Embeddings](#)
2. LII Legal Information Institute | [Wex Legal Dictionary and Encyclopedia](#)
3. European Commission | [DGT-Translation Memory](#)
4. ACL Anthology | [BLEU: A Method for Automatic Evaluation of Machine Translation](#)
5. ACL Anthology | [METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements](#)

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises, and communities to create value. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale, and build connected ecosystems. Connect with us at infosystopaz@infosys.com.

For more information, contact askus@infosys.com



© 2026 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.