Proceedings of the ASME 2026
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC-CIE2026
August 23–26, 2026, Houston, TX

DETC2026-194799

CAPTURING TACIT DESIGN FOR MANUFACTURING KNOWLEDGE FROM UNSTRUCTURED DISCOURSE

Douglas P. McGowan[1], Kosa Goucher-Lambert[1,*]

[1]University of California, Berkeley, Berkeley, CA

ABSTRACT

Implementing design for manufacturing in the early stages of product development can have significant effects on timeline and budget, but can be a difficult undertaking. Current approaches to automating rule-based feedback using large language models (LLMs) and machine learning do not adequately capture the complex manufacturing context around a product. A large part of the knowledge and experience used in giving manufacturability feedback is either tacit or lacking in structured documentation. This makes it difficult for designers to access this knowledge when seeking to consider manufacturability. This also presents a challenge for AI-in-design interventions, as there lacks an accurate way for AI models to access and reason with this tacit knowledge. Thus, we present a knowledge extraction method that enables AI to better augment early-stage concept design. Using post data from machining forums, we capture the tacit knowledge and experiences communicated by machinists and engineers, structuring it into a scalable knowledge base. The base consists of a multi-vector embedded space and a 3879-frame knowledge graph. Frame quality is validated using an automated LLM judge calibrated against a stratified 27-frame human-expert panel. This contributes a unique dataset on machining discourse and tacit knowledge, while demonstrating a method for transforming unstructured natural language data into useful design for manufacturing insights. These contributions enable AI to function as an extradisciplinary sensemaker that expands the designer's capacity to account for external constraints.

Keywords: design for manufacturing, knowledge graph, human-AI interaction

1. INTRODUCTION

In mechanical product design, accounting for manufacturing constraints is an essential part of the design process and has significant impacts on the later stages of product development. Design-for-manufacturing (DFM) is the processes and decisions in design focused on manufacturability, production cost, and defect risk. DFM is complex, and requires expert-level manufacturing knowledge that can be difficult or inefficient for designers to learn and access. This paper explores how DFM expertise can be extracted from unstructured natural language data and structured into an agentic reasoning-oriented knowledge base. This knowledge base can help future AI systems give more accurate and thorough design feedback grounded in concrete human expertise. We then test this knowledge base through an agentic DFM feedback workflow.

A large part of a product's lifecycle cost is committed during the early design stage [1, 2], and late-stage rebuilds risk overextending project timelines [3]. The concept generation and selection processes that occur in early-stage mechanical design could benefit from greater manufacturability awareness, but early-stage design for manufacturing (DFM) remains difficult [4, 5]. The rule-based geometric constraints common in most forms of DFM do not adequately capture the numerous considerations needed to design a manufacturable part.

Common examples of rule-based DFM heuristics include automated checking of problematic features like deep pockets, thin walls and small fillets. Digital tools like Siemens NX Checkmate, DFMPro and CoLab perform DFM analysis of CAD designs. These tools use geometry rules, historical quoting data, and supplier databases to evaluate a part's manufacturability.

The heuristics used by digital DFM platforms guard against common mechanical design pitfalls, but a large part of DFM failures remain unaddressed [5, 6]. Examples of these include failures caused by limitations in manufacturer capability, context-dependent process constraints, and assembly-level conflicts. Failures like these cannot simply be addressed with the available tools and processes [4, 5]. Many unsolved DFM roadblocks trace back to the tacit nature of manufacturing expertise. Machining is a clear example: seasoned practitioners hold deep experiential knowledge that is either undocumented or entirely tacit [7, 8]. This tacit and experiential knowledge plays a crucial role in the thought process behind manufacturability feedback, but is diffi-

*Corresponding author: kosa@berkeley.edu

cult for designers to know during early-stage design. Structuring tacit and experiential knowledge for digital communication is also challenging. While researching DFM, this paper focuses on machining as a subdiscipline due to its prevalence in multiple industries and the significant amount of unmapped expertise held by seasoned practitioners [9].

Current DFM research focuses on the automation of geometric assessment using computational tools and AI integrations [4, 6]. This can accelerate formal heuristics implementation, but is not effective for involving nuanced context-dependent information–manufacturer capabilities, equipment, and past experiences–that varies across organizations, settings and times. This information is essential for DFM and often drives design feedback that requires costly rebuilds. For manufacturers, communicating their requirements and constraints to designers early or clearly enough remains difficult because their feedback depends on tacit and context-dependent information and knowledge. Thus, there exists a need to capture experiential and tacit knowledge in an efficient manner.

There is also an opportunity to explore how to structure this knowledge for future AI-in-design interventions. Currently, AI tools like large language models (LLMs), vision language models (VLMs), and reasoning agents have demonstrated capability in providing feedback, performing research, and generating ideas in the design process. By structuring expert knowledge for agentic reasoning and retrieval, we can cultivate the expertise and contextual awareness of an AI model for DFM interventions to be more reliably accurate.

We explore the overarching topic: In what ways can knowledge graphs capture experiential knowledge in design and manufacturing from discourse and conversation data? In parallel, we focus on the following research questions:

1. What types of knowledge do technical practitioners communicate in online forums?

2. What types of knowledge and discourse can be captured in a reasoning-oriented format using knowledge graphs and embedding spaces?

3. How can subjective semantic review be automated using large language models (LLMs)?

To address these questions, we develop a multi-vector embedded space and a knowledge graph constructed from a corpus of design and machining forum posts. Online technical forums have been shown to be high-quality repositories of expert knowledge [8, 10], containing discourse with valuable practical insights that are not always captured in traditional textbooks. With the forum post data, we demonstrate a process for extracting specific pieces of knowledge from unstructured natural language, referred to as *frames* in this paper—drawing on Minsky's frame representation paradigm [11]. An automated review process calibrated on human expert feedback is used to verify the accuracy and utility of the frames, and an agentic application case study demonstrates the broader applications.

2. RELATED WORK

This section contains a review of relevant research. We start by establishing the limitations of current DFM processes and identifying tacit knowledge as a critical gap. Then, we survey existing methods of capturing knowledge and using LLMs to validate semantic accuracy. Finally, we discuss growth opportunities for AI's role in design.

2.1. Addressing the Tacit Knowledge Gap

Earlier research into this same knowledge gap highlighted key causes that remain today: lack of extra-disciplinary knowledge, organizational rigidity, and insufficient integration of manufacturing data during the design phase [2, 5]. Making this knowledge available to designers in a better way first requires capturing it. In general, there is a great part of knowledge and reasoning that can be classified as tacit, in that it is important and often referenced, but rarely explicitly stated and methodically documented. The communication gap between expert designer and expert manufacturer is multifaceted, but one impactful part of it is the gap in tacit knowledge.

The knowledge creation theory confirms the tacit nature of some parts of technical knowledge, showing the hands-on practice involved in acquiring it [7]. Both the learning and the documentation are informal and unstructured, which allows for agility in developing new practices, workarounds, and adjustments in a dynamic environment. However, this dynamic nature of the manufacturing environment also makes it difficult to standardize knowledge, which keeps much of tacit knowledge from becoming explicitly documented. There often is no official place for tacit knowledge to be discovered or referenced, except through extended communication and manual feedback. Furthermore, this tacit knowledge may also be lost. Recent research shows a risk of losing large amounts of tacit knowledge due to the aging workforce and transitioning industry requirements [8, 9]. If this knowledge is not captured, it may not be accessible in the future.

Even now, in the early stage of design, incorporating expert tacit knowledge is difficult, as there are often multiple different angles of expertise from which to consider each design. Furthermore, tacit knowledge is highly context-dependent, and it can be difficult to explain when the knowledge factors in to a decision. Current methods of using LLMs to broaden and accelerate research and feedback risk hallucination and overgeneralization because of the lack of structured tacit knowledge and the inefficiencies of reasoning over large texts [12, 13]. The work in this paper demonstrates a way to map high-context tacit knowledge in a format over which language models can effectively reason.

2.2. Computational Knowledge Capture Methods

Collecting reasoning-accessible knowledge is a challenge of its own. Earlier research presented natural language processing (NLP) techniques to construct knowledge graphs from structured documents [14]. Multiple agents would collaborate on the data analysis, which mimics the social nature of tacit knowledge transfer [15]. Other work focused on computational design applications, analyzing CAD files and engineering drawings to infer design decisions [16, 17].

These methods allow for more structured knowledge that AI can reference and reason over. In processing this data, LLMs read and structure the data, while also collecting data from workers directly. Related work also identifies the weakness of this process, as it only captures the structured parts of the design and manufacturing process, with limited exploration into the tacit, informal decision-making and experiential learning elements [15, 18].

A paper by Goridkov et al. conducted a study to collect data on the experiential knowledge collection process, mentioning future potential for scalability [19]. Other related work focuses on developing a multimodal dataset on design failures to form an interactive embedded space and a predictive model for potential design failures [20]. The embedded space was used to show designers related product failures during the ideation stage, and the LLM-powered predictive model was shown to reach near-expert reliability.

Collectively, this past research points to a gap in collecting, analyzing and structuring engineering knowledge. Our work demonstrates a streamlined method of capturing experiential knowledge by using existing forum data to capture both tacit and experiential knowledge. This also allows for collection of a larger and more diverse dataset, enabling future agentic systems to use a more versatile knowledge base. Our work expands on existing data structuring methods by combining an embedded space with a semantic knowledge graph, enabling a two-fold source of historical and experiential data. This format makes the dataset accessible through both manual searches and agentic retrieval.

2.3. Validating Data using LLMs

For large datasets like the one in this paper to be reliable for use in engineering design, it must be carefully validated. This is a demanding task for humans alone; recent research has investigated the use of LLMs to validate both designs and datasets. While human review involves having engineers with machining expertise review knowledge extractions or graph reasoning outputs for quality, the LLM-as-a-judge automates this process [21]. While 80% agreement can be found in general areas, this agreement does lessen when considering expert domains [22]. There is also a demonstrated bias toward lay user preferences due to biases in LLM training data [23]. This shows that, in niche domains, human review remains necessary to verify accuracy. However, it also verifies that, in niche domains, a general language model is insufficient to provide expert insight into a specific context, thereby confirming the need for more specific knowledge capture.

In disciplines where crowdsourcing expertise is difficult but expert review is still needed, LLM-as-a-judge methods have been developed to ensure the automated review is still valid [24, 25]. Multi-modal reasoning models are typically used to review designs, as the chain-of-thought reasoning has been shown to provide more reliable results [26]. Known weaknesses like flattery, position, and verbosity biases can be mitigated with prompt engineering and randomized ordering [27]. To verify the LLM's accuracy, expert-grounded measurements such as gold standard comparisons, inter-rater reliability, equivalence testing, and top-set overlap are essential [26].

Our work tackles a more complex task than previous research, as the data being reviewed is subjective and multifaceted. By testing multiple metadata and quality metrics, we gain insight on the ability of LLMs to review unstructured discourse data and understand semantic nuance. In domains where complex expert feedback is needed, LLM agreement has been shown to drop to 60–68% [22]; thus, human reviews are still essential for calibrating the automated review process. After the human-review validation step, automated quality control using LLMs is possible using gold standard and in-context learning techniques discussed in previous work. This validation process supports the use of this dataset in AI-powered design-for-manufacturing interventions within the machining domain captured by the corpus.

2.4. Towards AI for Knowledge Augmentation in Design

Existing research on AI in design shows that AI functioning in a feedback-providing role gives the designer agency that translates to better outcomes [28]. While much research has been conducted on AI's role in design and concept generation, there remains a significant opportunity for AI to grow in its role as a cognitive augmenter that enhances the designer's decision making [29, 30].

Across the discussed research, no existing approach structures tacit manufacturing knowledge at scale from informal discourse, structures it for machine reasoning, and validates it through both expert and automated review. By adapting a computational knowledge-capture methodology to a high-demand discipline of design for manufacturing, this work contributes both a valuable knowledge base and structuring methodology. The applications and utility of this work are made clear by the recent research demonstrating the validity of LLM-as-a-judge and AI's role as a knowledge architect in product design [31, 32].

3. METHODOLOGY

This section describes a four-stage pipeline for transforming unstructured online forum discourse into a structured, machine-queryable DFM knowledge base: (1) data collection and filtering, (2) knowledge frame extraction, (3) multi-vector embedding, and (4) knowledge graph construction. The pipeline, shown in Figure 1 preserves the nuance of forum discourse while producing structured knowledge representations that an AI agent can reason over, in real time, during design tasks.

3.1. Data Collection and Filtering

To collect the data, we scraped forum threads from Stack Exchange. Online forums have been shown to contain useful practitioner knowledge, with experienced users answering the majority of posted questions. The conversational nature of forums allows for more nuanced and context-dependent knowledge to be shared [33, 34]. Forum data was collected via the Stack Exchange API under the platform's Creative Commons Attribution–ShareAlike (CC BY-SA 4.0) license, which permits reuse with attribution.[1] No private or user-identifying information was collected beyond publicly displayed usernames and

[1] See https://stackoverflow.com/help/licensing and https://creativecommons.org/licenses/by-sa/4.0/.

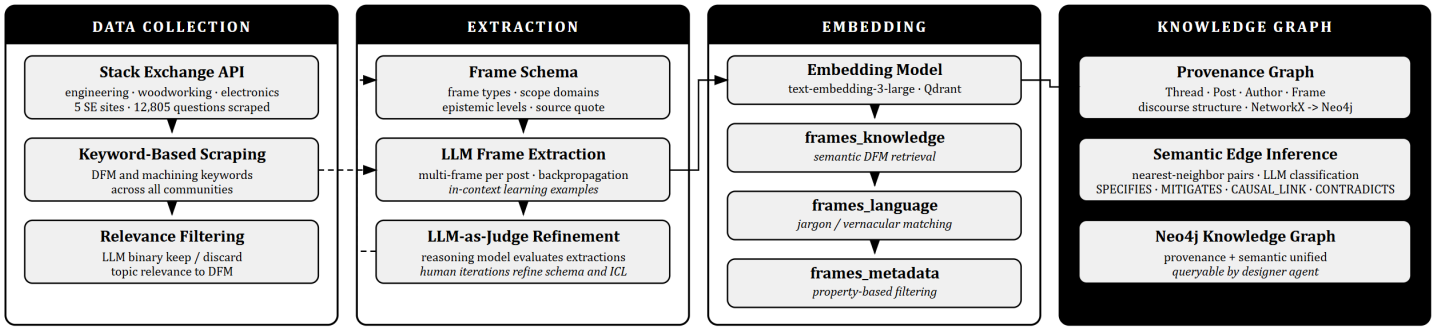

FIGURE 1: Pipeline overview for extracting and structuring machinist knowledge into a queryable framework.

reputation scores. Since there are a wide variety of conversations held on Stack Exchange, we developed a system to find the most relevant discussion threads. To capture both design-oriented posts and machining-process posts with potential design implications, a list of approximately 20 keywords was developed spanning two complementary categories: keywords describing design decisions and constraints, and keywords describing machining problems or failure modes.

We ran searches with these keywords across five different Stack Exchange sites: Engineering, Mechanics, Woodworking, Electronics, and Robotics. With the Stack Exchange API, we scraped 12,805 raw threads, each containing multiple posts. Using GPT-5-mini as a relevance filter, we removed questions that lacked relevant content, such as off-topic or non-technical discussion, and distilled the dataset to 1,936 posts. GPT-5-mini was selected for the filtering task because its lower latency and cost are well-suited to high-volume single-decision classification calls; recent LLM-as-judge surveys document that smaller models perform comparably to frontier models on low-complexity classification tasks such as topical relevance gating where the decision boundary is well-defined [24, 25]. Full quantitative metrics on the dataset and its pipeline, including the filter's per-site keep rates and prompt structure, can be found in Appendix A, particularly §A.1.

3.2. Knowledge Frame Extraction

To create the knowledge base, we developed a framework for extracting knowledge from the forum data. Raw forum posts have valuable lessons and insight, but in a format that is difficult for a machine to reason with. There is informal language, narrative structure, implicit assumptions, and high-context claims.

To take advantage of this richness in discourse, we introduce the concept of a knowledge frame—drawing on Minsky's frame representation paradigm [11]—defined as one distinct unit of design or manufacturing knowledge that is extracted from a post by an LLM with reasoning capabilities. Each post often contains knowledge from which multiple different frames can be extracted. For example, a machinist recounting a story of a build failure could include universal lessons, case-based observations, and clever workarounds in one post.

A 23-field schema was developed for each frame. We developed the schema iteratively based on existing research on structuring knowledge [35–37], with a focus on preserving nuance by separating the different semantic qualities of each knowledge claim. The schema served to both constrain the LLM's extraction output and structure the required metadata for future analysis. The fields were intentionally structured to capture various knowledge facets. Table 1 below shows the full schema structure.

The frame_type classifies what type of claim the frame is making, while the post_role classifies the post's main purpose in the discussion and the epistemic_stance classifies how the post author views the claim being made. These fields help the LLM classify the frame and guide the main_point field, which captures the crux of the frame. The scope, subject, and applicability fields force the LLM to specify where the frame is useful. To force a separation between hypothetical values and discrete heuristics, the example_value and example_context fields have targeted prompts. Finally, the extraction_rationale contains any extra reasoning for the LLM to explain a potentially confusing frame classification.

When prompting the LLM, we used a few-shot approach, embedding examples for each schema field individually. Our prompt encouraged the LLM to extract multiple frames per post to prevent the LLM from collapsing a multi-faceted statement into only a single frame. After this, there was a multi-part filtering step.

The first filtering step used few-shot prompting to check the epistemic_stance field. This field is particularly important for categorizing the credibility and nature of the frame, and additional steps were required to ensure that the nuance of implicit communication was accurately captured. The second filtering step analyzed the source_quote field to determine whether duplicate frames were extracted. Any frames with similar quote citations were reanalyzed, with the LLM deciding whether to merge the frames, keep both frames, or delete one of them. Any frames with inaccurate source_quotes were deleted, and if a post was flagged for relevant content but no frames were extracted, the system was instructed to retry extraction.

Before using the schema for full frame extraction, we iterated on the schema and extraction scripts. Out of a total 192 iterations, there were 56 human iterations and 136 LLM iterations on the code set. For the human iterations, we manually reviewed the extraction results from a set of diverse test frames. Based on the results, we adjusted the schema and extraction script to better capture the knowledge in the post. For the LLM iterations, we applied LLM-as-a-judge evaluation methods for each iteration. We used GPT-5 to assess the schema and extraction script alongside the post and its extractions, and instructed it to

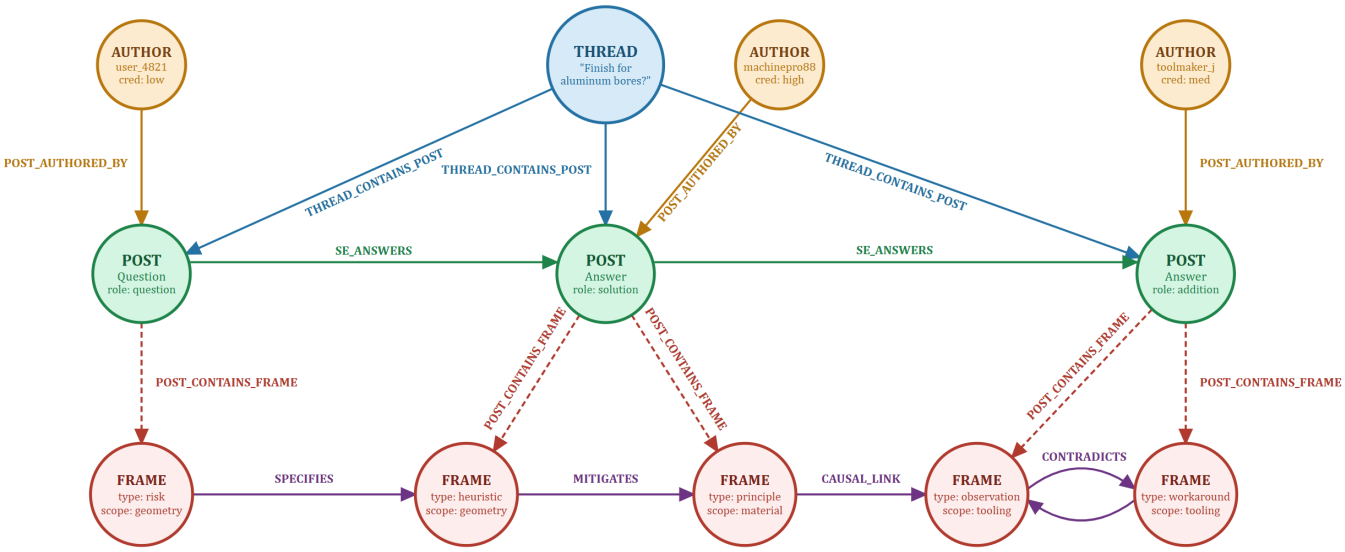

FIGURE 2: Knowledge graph architecture flowchart.

make direct, surgical changes to the scripts before iterating again. GPT-5 was chosen as the main extraction and graph-construction model because, at the time of the build, it was the state-of-the-art reasoning-capable model from OpenAI within the cost/latency band suitable for the corpus scale [38].

The LLM-as-a-judge iteration involved in-context learning (ICL) based on gold-standard examples, which helped schema convergence. After converging on a schema that could successfully extract frames from posts in multiple different design or manufacturing disciplines, we executed the full extraction. Out of the 1,936 posts, 1,203 posts had relevant content and 3,879 frames were extracted, with an average of 3.22 frames per post.

3.3. Validation

To ensure quality of frames, we used an automated LLM pipeline to review the frames' quality and revise them as necessary. Following the methods discussed in Section 2, we conducted human review of the frames, with three engineering experts reviewing 27 stratified-sampled frames. The frames were reviewed for multiple metrics, including clarity, accuracy, and applicability. The reviewers also rated how tacit the extracted knowledge was, which is inversely related to how easily the knowledge could be found through simply reading a textbook. Each frame was given nine quantitative ratings and one qualitative rating (keep/revise/remove). Given the complex context of each frame and its rating, fewer frames were selected to minimize reviewer fatigue and meet scheduling constraints.

Once this gold standard was established, we created a rubric-based prompting structure to have GPT-5 review each frame and apply the qualitative rating. This forced the LLM to decompose the evaluation into discrete criteria. Using the gold standard, we implemented in-context learning. Once each frame had a rating, the next prompting script executed each command, revising and removing the required frames. This separation of the judgement and generation steps in the LLM aligns with recommendations from existing research [21, 25]. Since reviewing natural language knowledge extraction data is a subjective task, especially for tacit knowledge, we focused on the overall levels of harshness exhibited by the human reviews. Harsher reviews required more textbook-like and universal principles, and lenient reviews were more flexible with anecdotal and context-specific insights. Using the human review data, we calibrated the LLM's harshness and grading habits to match that of the experts.

These validation steps improve the reasoning-oriented reliability of the dataset as a retrieval-grade representation. Classification-label dimensions with lower IRR (frame_type AC1 = 0.50, applicability AC1 = 0.58) are presented as advisory metadata rather than authoritative filters; see Section 5.3.

3.4. Multi-Vector Embedding

Once the frames were extracted, a 3-vector embedded space was constructed using OpenAI's `text-embedding-3-large` model (3,072 dimensions). This model was selected because, at the time of construction, it was the highest-performing general-purpose embedding model on the public MTEB benchmark suite within the required size for our corpus scale (3,879 frames $\times$ 3 vectors) [39]. The first vector embeds the extracted knowledge content, pulling from the subject, main point, example context, example value and thread context. This supports semantic retrieval for DFM knowledge and advice, enabling the most relevant knowledge to be found, independent of how it was communicated by the author of the original post. The second vector was based entirely on the source quote, to capture the semantic meaning of the natural language used by the author based on linguistic similarity. This helps with queries phrased with jargon or informal speech. The third vector is an auxiliary metadata vector, based on the metadata rendered as natural language which includes: frame type, scope, epistemic stance, and post role. This supports property-based querying of frames, especially queries based on knowledge type.

A script that supports single-vector and multi-vector search was built and used to test this embedded space. Alongside this,

TABLE 1: KnowledgeFrame Schema Fields

Section	Field	Purpose
Post Metadata	`author_username`	Display name of the post author on Stack Exchange
	`post_id`	Unique identifier for the post from the normalizer
	`post_date`	ISO date string of when the post was created
	`post_role`	Discourse role of the post in the thread (e.g. question, solution, addition, war story)
	`post_score`	Net votes (upvotes minus downvotes) from the platform
	`is_accepted_answer`	Whether this post was marked as the accepted answer
	`question_id_ref`	Parent question post ID, for answer-sourced frames
	`source_url`	URL to the original Stack Exchange question page
Knowledge Content	`source_quote`	Verbatim substring from the post supporting this claim; provenance anchor
	`main_point`	One transferable knowledge claim in 2–3 sentences, fully supported by the source quote
	`subject`	Concise noun phrase for the primary concept; serves as the graph node anchor
	`example_value`	Concrete scenario or worked example from the post that illustrates the claim
	`example_context`	Real-world context scoping the claim; required when applicability is not universal
	`extraction_rationale`	LLM reasoning for extraction decisions; required when frame type or scope is `other`
Frame Metadata	`frame_type`	Epistemic category: `risk`, `heuristic`, `principle`, `observation`, `comparison`, `workaround`, `case`, or `other`
	`scope`	Knowledge domain: `design_geometry`, `machining_process`, `material`, `tooling`, `programming`, `shop_operations`, `quality`, `cost`, or `machine_capability`
	`applicability`	Breadth of the claim within its subject: `universal`, `context_dependent`, or `situational`
	`epistemic_stance`	How the claim is expressed: `absolute_authoritative`, `direct_neutral`, `hedged_emphatic`, `tentative_subjective`, or `narrative_implied`
Human Review	`validated`	Whether a human reviewer has confirmed this frame
	`validation_notes`	Reviewer notes: edits made, contradiction context, follow-up flags
	`contradicted_by`	Frame IDs of frames that contradict this one
	`related_frames`	IDs of related or supporting frames
	`thread_context`	Summary of the thread context surrounding this frame

we used topic modeling with BERTopic [40] to surface overarching themes and recurring subjects from the data. Using the existing embedding space, we used a UMAP [41] reduction to reduce the 3,072-dimension vector space to 5 dimensions, with $n\_neighbors = 15$. In this 5-dimension space, we used HDBSCAN [42] to find 319 topic clusters based on density. These topics were reduced to 39 larger clusters using BERTopic's hierarchical topic reduction. For easier visualization for results, these 39 topics were grouped into 8 high-level groups. These groupings could then be compared with the predetermined scope definitions set for the frames for further analysis.

3.5. Knowledge Graph Construction

Finally, a knowledge graph was constructed to capture the structure of the online discourse and the semantic relations between related frames. At first, we built a provenance graph with AUTHOR, THREAD, POST and FRAME as node types, and THREAD_CONTAINS_POST, POST_AUTHORED_BY, POST_CONTAINS_FRAME, and SE_ANSWERS as edge types. This part of the graph captures the discourse structure and shows human connections between different threads.

To construct the semantic portion of the graph, we found the top 5 nearest neighbors of each frame in the embedded space (based on all 3 vectors). For each unique pair found, GPT-5 inferred semantic relationships between the frames that had relevant content [13]: SPECIFIES, MITIGATES, CAUSAL_LINK, and CONTRADICTS. Unrelated node pairs had no edge drawn. This inference process added another 4,372 semantic edges. Figure 2 illustrates how the nodes and edges connect structurally and semantically. All nodes and edges were merged into a unified Neo4j knowledge graph. Full quantitative metrics on the human-expert panel, the GPT-5 LLM-judge full-corpus review, and the pipeline ablation study are reported in Appendix C.

4. RESULTS AND EVALUATION

This section presents the results of the knowledge extraction process, both through quantitative metrics and qualitative discussion. We first discuss the data validation results. This is followed by a discussion of the metadata, primary data outputs, and opportunities for future work.

4.1. Dataset Characteristics

To address RQ1, we examine the distribution of knowledge types across the extracted frames. Across the different frame types and topics, the vast majority of the frames were heuristic (1401/3879) and observation (1345/3879). This aligns with the typical function of forum posts, in which users focus on providing advice, solving problems, or identifying relevant information [10]. By predetermined scope, the largest category was design geometry, which confirms the large presence of useful design information in engineering forums. Besides the "other" category, machining process and machine capability were the second and third largest categories, respectively. Full details of the frame metadata distributions are provided in Appendix A.

Our first output from this frame dataset was a multi-vector embedded space. The UMAP visualization, shown in Figure 3, demonstrates the difference between the predetermined and clustered topic models. The clustered topics show a clearer distinction along zones of semantic similarity, while the predetermined topics are more intermixed. The predetermined topics, while an example of common fields of consideration in DFM, do not effectively organize the discussion threads and frames as the topic clusters do. An example of this is the tooling and machining process topics, whose centroids on the graph are relatively close. This

pattern is understandable, as manufacturing problems and solutions often involve multiple subdisciplines. The clustered topics focus more on specific activities or skills, which differs from the predetermined topics' emphasis on areas of analysis. While textbooks and reports may organize information more similarly to the predetermined topics, an adaptive classification approach is better for organization when looking at unstructured data.

Our second output was a knowledge graph, fully visualized in Appendix A. This contains the full structure of both the forum discourse and the semantic connections between frames. The majority of the nodes in the graph are connected, but there are several isolated groups of nodes with no connections to the central cluster. These signify discussions that are within the discipline of design for machining, but are not connected to the larger body of discussion. This illustrates the diverse nature of forum posts, where multiple unrelated discussions about similar topics are hosted on the same platform. The benefit of combining the knowledge graph with the embedded space is that, even if the user navigates a smaller, disconnected cluster, they can still search a wider area in the embedded space. This also works with AI agents, which can traverse the graph and the embedded space to retrieve the most relevant knowledge. These results address RQ2: the combination of multi-vector embedding and a semantic knowledge graph preserves both the topical content and the relational structure of forum discourse. The three embedding vectors capture different facets (knowledge content, linguistic style, metadata), while the four semantic edge types (SPECIFIES, MITIGATES, CAUSAL_LINK, CONTRADICTS) provide an LLM-inferred relational layer for downstream agentic reasoning; edge-level human validation is identified as future work in §5.3.

4.2. Knowledge Frame Validation

The human review statistics of the 27 unrevised, stratified-sampled frames are presented in Tables 2 and 3. Due to the subjectivity of rating extracted knowledge, the inter-rater reliability (IRR) varied between metrics. This is understandable, as each human rater approached the same knowledge from a different subjective paradigm on design and manufacturing. However, the average scores on quality, clarity, and accuracy demonstrate the utility of the presented knowledge extraction strategy. Despite disagreements on the specific qualities that made the knowledge useful, the ratings show that each reviewer believed the knowledge was useful overall.

For each of the 27 frames, there were ten ratings given by the reviewers. The metrics on extraction accuracy, clarity, and main point quality focus on the overall validity of the frame; these metrics were used to assess the frames' IRR. The metrics on scope, post role, epistemic stance, and applicability provide insight on the LLM's agreement with humans in classifying natural language data. In addition to this insight, these metadata metrics were used to calibrate the automated frame review and revision pipeline. Each of these metrics contained "poor", "acceptable", and "good" tiers. The tacit knowledge rating asked users to classify whether the knowledge communicated in the claim could be found in a traditional textbook. After the first 9 qualitative ratings, the final rating for each frame asked the reviewers to recommend keeping, revising or removing each frame.

In analyzing the IRR of this grading scheme, we used Gwet's AC1 [43] to measure agreement in place of the alpha and kappa metrics used in past papers. Because of the vast majority of ratings being in the "good" category, the kappa and alpha ratings understate agreement—a well-documented statistical artifact known as the kappa paradox [44]. The perceived randomness of agreement in a rating where one category dominates makes the high agreement of this data seem by chance. For example, since 83–100% of ratings were viewed as "good", the 96% exact agreement on main point quality is actually shown as negative by the Fleiss kappa rating. However, the only way to get a higher rating for that measurement would have been perfect agreement. The three ratings used to measure frame quality had high agreement and high average quality ratings, seen in Table 2:

TABLE 2: Frame quality ratings across human reviewers.

Dimension	Good	Accept	Unsat	Agree	AC1
Extraction Accuracy	92%	8%	0%	84%	0.840
Clarity	92%	6%	1%	85%	0.846
Main Point Quality	99%	1%	0%	96%	0.962

Overall, the majority of the frames were rated highly, with a high percentage agreement and AC1 metrics. The keep/revise/remove ratings were also favorable, with 0 out of 27 frames being voted for removal, and 2 frames being recommended for revision by majority vote. Of the "revise" rated frames, the primary reasons were lack in clarity, and disagreements about metadata. When extracting frames, the LLM had the freedom to reference information from the overall post or original thread question, but was instructed only to reference specific quotes. After reviewing each of the "revise" rated frames criticized for clarity, we noticed that each of the frames were accurate, but referenced data outside the quoted post excerpt.

The metadata ratings, shown in Table 3, were more contested, with lower agreement and average ratings. There was significant disagreement among reviewers on the accuracy of each metadata metric, showing the subjectivity of the classification method. While these metadata fields are not essential to the dataset, they can provide a helpful way to search and organize the frames. This insight informed our structuring of the dataset in the case study discussed in Section 4.3.

TABLE 3: Metadata classification ratings across human reviewers.

Dimension	Good	Accept	Unsat	Agree	AC1
Scope	90%	9%	1%	81%	0.808
Post Role	88%	8%	4%	73%	0.712
Epistemic Stance	88%	5%	6%	69%	0.673
Applicability	86%	10%	4%	58%	0.577
Frame Type	77%	18%	5%	50%	0.500

The tacit knowledge rating showed an average of 71% of frames to contain tacit or expert-level knowledge, with strong disagreement among reviewers about which frames were truly non-textbook. This confirms the nature of tacit and expert-level engineering knowledge—the perception and possession of

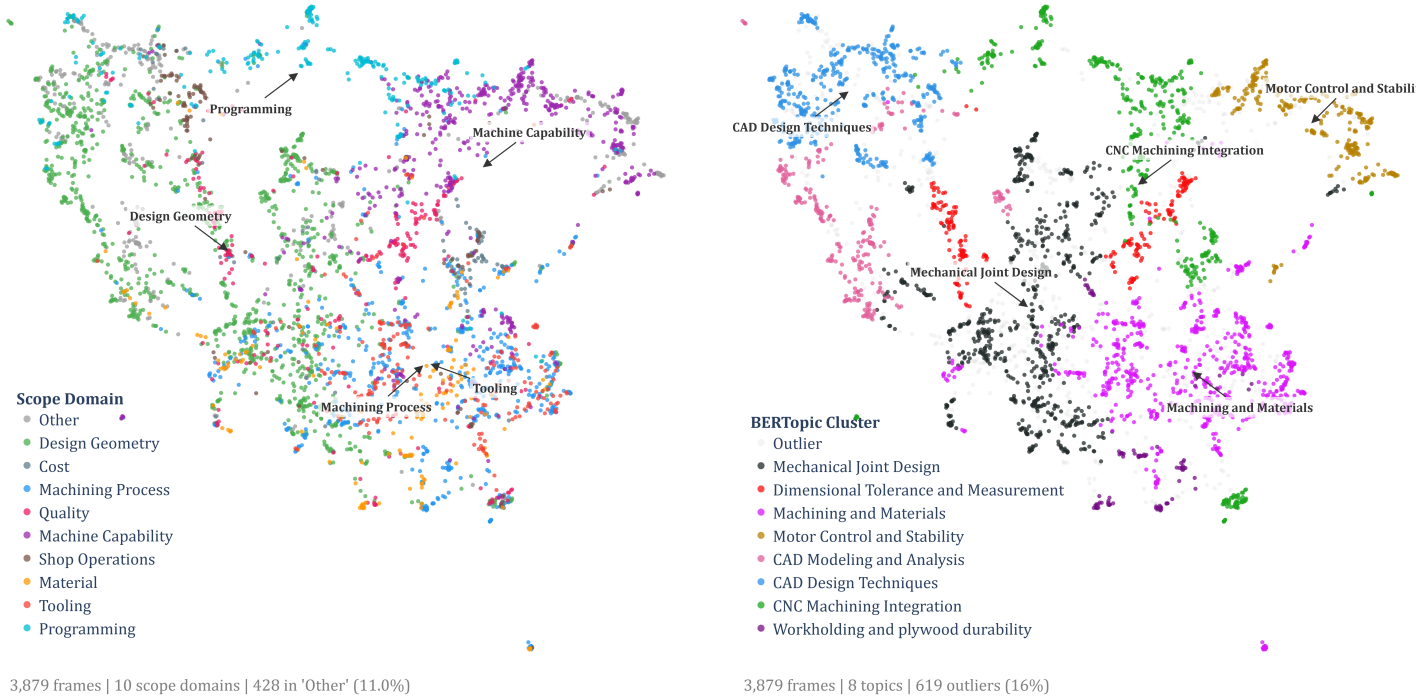

FIGURE 3: UMAP projections of the embedded space, shown as two complementary views. Left: colored by predetermined scope categories. Right: colored by BERTopic clusters.

it changes across different experts and sources [7, 8]. The tacit knowledge rating also showed that, while a significant portion of the extracted knowledge was tacit, there is still a strong presence of textbook information communicated in the forums.

Overall, the LLM was more conservative on content and tacit knowledge rating, but more generous on the metadata ratings. In general, we found that the LLM was a harsher critic of the quality of knowledge extracted. The LLM only found more agreement with human ratings once the gold standard demonstrating the tacit knowledge's usefulness was implemented. This produces an insight that LLMs might naturally discount tacit and experiential knowledge in favor of more structured or formal knowledge [22, 23]. Similarly, the lower tacit_knowledge rating given by the LLM shows that it tends to default to textbook knowledge for concepts that human reviewers might view as more tacit.

The expert review of the frames provided helpful calibration for the automated LLM rating and revision pipeline. Of the 3,879 frames reviewed, 20% were marked for revision, with only 3 frames marked for removal. To address RQ3, we find that automated LLM review is feasible for frame quality assessment (AC1 $\geq$ 0.84 on quality metrics) but requires human calibration for subjective dimensions like tacit knowledge classification. In total, the knowledge extraction process resulted in 3,879 frames, which created a corresponding embedded space and semantic knowledge graph with 4,372 semantic edges. The total graph with both semantic connections and discourse provenance contained 7,864 nodes and 13,148 edges.

Following best practices in evaluation methodology for LLM-as-judge studies in expert domains [26, 27], we report complementary statistics alongside AC1: per-frame all-3-rater exact agreement, and pairwise percent agreement averaged across all rater pairs. Reporting these alongside AC1 lets readers distinguish strong consensus from chance-corrected statistical agreement on noisy dimensions. We additionally report a GPT-5-judge repeatability check on a 10-frame subsample under fixed seed and prompt (Appendix B), which isolates judge stochasticity from cross-judge variation.

4.3. Case Study: Agentic Retrieval

To demonstrate the usefulness of this knowledge, we use the frame embeddings in a retrieval system. By constructing an agentic retrieval-augmented generation (RAG) system [12], we enable a more concrete connection between design advice and experiential commentary. In response to design queries, this agent retrieves vectors, traverses the knowledge graph, synthesizes a response, and identifies gaps in the knowledge base.

When making a query, the user can adjust three parameters to adjust the scope of knowledge retrieval: N, M, and Q. With every query, the model retrieves the N nearest frames, retrieves any semantically connected nodes within M degrees of removal, and down-selects to Q total frames. Based on these frames, it synthesizes a query response, detailed in Table 4, and identifies gaps where the available frames do not provide coverage. The model defaults to a concise response, only listing the reference frame IDs, but the user can easily access details on the frame citations.

If the user doubts the advice, they can request the context of the discourse behind the frames. In response to this query, the model accesses the provenance graph, and retrieves the structure of the discussions behind each frame, including the author names, author credibility scores, and complete thread content for each of the frames used as evidence. The agent then uses this data to

provide a summary of the existing context.

If the user requires information not covered in the available frames, they can request for the model to backfill the identified knowledge gaps with information from textbooks or official standard codes. This application demonstrates that datasets like this can augment the existing capabilities of language models to search for information with specific data on a complex discipline [13, 45]. When compared to a typical LLM response to the same design question, this agent provides a more detailed and grounded response, where there is a clear human link to each response provided. Furthermore, the textbook answers being incorporated into the knowledge gaps enable a comprehensive response. Using agentic retrieval, the best aspects of both general LLM searches and specific graph queries can be combined to provide useful insight. More details on the agent's reasoning process and output can be found in Appendix D.

TABLE 4: Structure of a synthesized DFM agent response.

Section	Description
Summary	Concise overview of DFM advice from all retrieved frames.
Discipline Sections	Categorizes retrieved knowledge into specific engineering domains (e.g., machining, materials).
→ Evidence	Sourced claims citing specific frame IDs.
→ Reasoning Chains	Sequences of frames linked by semantic edges (CAUSAL, MITIGATES, SPECIFIES).
→ First Principles	Governing physics, scaling rules, and engineering rationale.
→ Formulas	Relevant equations with variable definitions and notes.
Design Rules	Actionable guidance for the designer at the CAD/drawing stage, citing frame IDs.
Contradictions	Cases where two frames provide opposing guidance, presented without resolution.
Knowledge Gaps	Specific aspects of the user query not covered by the retrieved frames.
Gap Backfill	Textbook-level answers for gaps, with confidence notes and standards references.

To isolate the contribution of each pipeline component, we ran a three-way ablation against eight design questions spanning three coverage tiers (three well-covered, three sparse, two intentionally out-of-scope). Configuration A used GPT-5 with no retrieval; Configuration B added top-12 vector retrieval over the frames collection but no graph traversal; Configuration C used the full pipeline described above (decomposition, multi-vector retrieval, two-hop graph traversal, structured synthesis with a knowledge-gap channel). All three configurations emitted a common JSON output schema with capped rule/gap counts and a 200-word summary cap, to control for the verbosity confound that would otherwise inflate any "richer answer" metric. The load-bearing finding: only Configuration C correctly refused both out-of-scope queries (2/2 vs. 0/2 vs. 0/2), using the structured `knowledge_gaps` field to report that the corpus contains no evidence on the subject. Full per-query metrics (latency, cost, refusal flag, frame-ID faithfulness) and the verbosity-matched re-run protocol are reported in Appendix C.

5. DISCUSSION

5.1. Contextualizing the Contributions

This paper contributes a dataset with diverse knowledge claims and a reasoning-prone structure. The method of knowledge collection is a secondary contribution in that it explores the structuring of insight from unstructured discourse data. In collecting this data, we found that reviewing the quality of tacit and experiential knowledge extraction is a more subjective task than other common applications of automated LLM review [27, 30, 31]. To overcome this, we calibrate our automated review using similar techniques to past work, but with a different measurement of IRR and a more multifaceted rubric.

Similar work on automated review in expert domains shows 60-68% review agreement; while the ratings on frame quality exceed this, ratings on the metadata are both above and below this range [22]. This shows that assessing the usefulness of tacit and experiential insights is a more reliable task than the classification of those insights. This is especially evident when reviewing how tacit a knowledge frame is; the boundary between textbook and tacit depends on the education and experiential knowledge of the reviewer. Additionally, the LLM defaulted to discounting experiential knowledge in favor of more structured or formal representations.

Comparable work on distilling design knowledge relied on structured documents — product teardowns, CAD files, patents, and product recalls [14, 17, 19, 20]. Forum posts surface knowledge types absent from those sources: observations, workarounds, heuristics, and practitioner contradictions, at a scale that expands on prior experiential-capture efforts [19]. Topic modeling revealed that activity- and skill-focused clusters organize discourse more naturally than predetermined analytical categories. The graph's disconnected nodes confirm that RAG and graph traversal must be combined to avoid missing relevant but isolated frames.

The dataset's topic distribution shows that design geometry, machining processes and machine capabilities are the topics with the highest density in questions asked, and that practical heuristics and empirical observations are the most common forms of knowledge shared in these forums. This helps the DFM process in a way that traditional rule-based processes do not. Furthermore, the CONTRADICTS semantic edges provide a look into areas where practitioners disagree, which also provides nuance that traditional textbooks and automated DFM checks do not. Finally, the graph's structure links each knowledge frame to a specific source post and author, supporting provenance audit by downstream readers.

5.2. Opportunities for AI Knowledge Augmentation in Design

For designers, this work shows that design feedback based on human expertise and experience can be retrieved computationally. Designers can find feedback that is both detailed and practical before consulting an external manufacturing expert, whether through manual search or agentic retrieval. In complex technical fields such as design, experts make decisions and generate ideas using a large internal base of tacit and experiential knowledge. As

shown by this dataset, much of that knowledge is context dependent, and some of it is contested between different perspectives. AI can provide assistance in navigating and understanding this knowledge and has been shown to be useful in generating feasible but less diverse concepts [46]. Nonetheless, maintaining agency in the designer is important as intelligent digital tools continue to grow in number and capability. Thus, a paradigm where AI is a sensemaker that surfaces and interprets knowledge for the user allows for more constraint-aware technical decision making and idea generation.

An intentional choice in this work is that the knowledge graph and example agentic applications do not generate concepts. The knowledge graph helps the agent surface relevant knowledge in response to queries, preserving the designer's role in creating and refining concepts. Recent empirical research shows that generative AI in concept design can result in premature convergence compared to human-only teams, where fewer alternatives are explored and functional quality is lower [32]. Other work shows that access to generative AI increases individual creativity at the cost of collective output diversity [47]. Thus, our work positions AI not as a generator, but as a collaborative advisor that surfaces experiential knowledge, flags contradictions between practitioners, and identifies gaps where no knowledge exists, while leaving creation to the designer.

A recent review draws parallels between human working memory limitations and LLM context windows [48], showing that retrieval-augmented generation extends an LLM's effective knowledge in the same way that consulting an external expert extends a designer's effective knowledge. Agentic systems that perform research through techniques like RAG for the designer can enable multidisciplinary design perspectives while reducing the designer's cognitive load. This is consistent with findings that AI assistance maximizes value when task complexity exceeds the capacity of a single designer [49].

5.3. Limitations and Future Work

This paper demonstrates the ability to extract structured knowledge from unstructured data. However, using forum posts as the only data source for the knowledge base provides some limitations. Expertise is not required to be active and influential on public online forums, which means claims made can vary greatly in validity and technical accuracy. In addition, the use of LLMs to extract knowledge frames provides another source of potential error, as reasoning models like GPT-5 do not always produce repeatable results. This was seen most in the metadata classifications.

Future work could improve these areas by combining extractions from multiple sources of data. Data from multiple discussion platforms can be used to expand the knowledge graph to enrich and broaden the knowledge base. In addition, augmenting this dataset with frames extracted from machining reports or manuals would allow for another layer of contextual knowledge in the graph. Even comparing extraction quality between models from different companies could provide insight into how LLM architecture affects automated analysis tasks.

These changes would enable future applications of the work to be more accurate and versatile. The presented examples of knowledge retrieval demonstrate the dataset's potential use, but real-world validation remains an open opportunity. By assessing changes in design outcomes when using this tool in academia and industry, information on the dataset's true utility can be found. Moreover, the effects of this automated DFM feedback on concept generation and selection could be analyzed. A study could examine the outcomes of both divergent and convergent concept design sessions with DFM intervention.

5.3.1. Expanded Applications The pipeline generalizes to any domain where expert knowledge is largely tacit, discourse data exists, and a large practitioner population requires that knowledge. In engineering, FEA and physical testing are clear candidates. FEA requires frequent judgment calls on mesh density, boundary conditions, and result interpretation that senior analysts accumulate through experience but rarely document; a reasoning-oriented knowledge base could surface those modeling heuristics directly. Physical testing presents a similar pattern, where decisions about fixture design, sensor placement, and anomaly interpretation are typically transferred through mentorship rather than formal documentation [7, 8]. The three-part criterion — tacit knowledge, available discourse, large practitioner base — also maps onto fields outside engineering entirely. Healthcare and legal practice both involve substantial bodies of experiential knowledge embedded in informal discourse [45, 50], though patient privacy, attorney-client privilege, and domain-specific regulatory requirements introduce constraints that future work must address before the pipeline could be responsibly applied.

6. CONCLUSION

This paper demonstrated a method for extracting, structuring, and validating tacit machining knowledge from online forum discourse. From 1,936 posts by 1,138 authors, we extracted 3,879 knowledge frames organized into a multi-vector embedded space and a semantic knowledge graph. Human expert review on a 27-frame stratified panel confirmed extraction quality on the dimensions where inter-rater reliability supports authoritative use. An LLM-as-a-judge pipeline calibrated against these expert ratings then extended quality control across the full 3,879-frame corpus. The agentic retrieval case study showed that structured tacit knowledge can augment LLM capabilities for design-for-manufacturing feedback [12, 13]. By grounding AI in domain-specific experiential knowledge rather than general training data, this work contributes to a paradigm where AI expands the designer's cognitive bandwidth while preserving their creative agency [28, 29].

ACKNOWLEDGMENTS

This work has been supported by the Regents of the University of California. The findings presented in this work represent the views of the authors and not necessarily those of the sponsors.

REFERENCES

[1] Ehrlenspiel, Klaus, Kiewert, Alfons and Lindemann, Udo. *Cost-Efficient Design*. ASME Press / Springer (2007). DOI 10.1007/978-3-540-34648-7.

[2] Boothroyd, Geoffrey, Dewhurst, Peter and Knight, Winston A. *Product Design for Manufacture and Assembly*. Marcel Dekker (1994).

[3] Chang, Tien-Chien, Wysk, Richard A. and Wang, Hsu-Pin. *Computer-Aided Manufacturing*, 3rd ed. Prentice Hall (2006).

[4] O'Driscoll, Michael. "Design for Manufacture." *Journal of Materials Processing Technology* Vol. 122 No. 2–3 (2002): pp. 318–321. DOI 10.1016/S0924-0136(01)01132-3.

[5] Gupta, Satyandra K., Regli, William C., Das, Diganta and Nau, Dana S. "Automated Manufacturability Analysis: A Survey." *Research in Engineering Design* Vol. 9 No. 3 (1997): pp. 168–190. DOI 10.1007/BF01596601.

[6] Formentini, Giovanni, Rodriguez, Nicolas B. and Favi, Claudio. "Design for Manufacturing and Assembly Methods in the Product Development Process of Mechanical Products: A Systematic Literature Review." *International Journal of Advanced Manufacturing Technology* Vol. 120 (2022): pp. 4307–4334. DOI 10.1007/s00170-022-08837-6.

[7] Nonaka, Ikujiro and Takeuchi, Hirotaka. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press (1995).

[8] Sun, Xin, Huang, Rui, Jiang, Zheng, Lu, Jian and Yang, Shuai. "On Tacit Knowledge Management in Product Design: Status, Challenges, and Trends." *Journal of Engineering Design* Vol. 35 No. 1 (2024). DOI 10.1080/09544828.2023.2301232.

[9] Deloitte and The Manufacturing Institute. "Creating Pathways for Tomorrow's Workforce Today: Beyond Reskilling in Manufacturing." https://www2.deloitte.com/us/en/insights/industry/manufacturing/manufacturing-industry-diversity.html (2024). Industry report.

[10] Mamykina, Lena, Manoim, Bella, Mittal, Manas, Hripcsak, George and Hartmann, Björn. "Design Lessons from the Fastest Q&A Site in the West." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*: pp. 2857–2866. 2011. ACM. DOI 10.1145/1978942.1979366.

[11] Minsky, Marvin. "A Framework for Representing Knowledge." Winston, Patrick Henry (ed.). *The Psychology of Computer Vision*. McGraw-Hill (1975): pp. 211–277.

[12] Lewis, Patrick et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems*. 2020. URL https://arxiv.org/abs/2005.11401.

[13] Pan, Shirui, Luo, Linhao, Wang, Yufei, Chen, Chen, Wang, Jiapu and Wu, Xindong. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." *IEEE Transactions on Knowledge and Data Engineering* Vol. 36 No. 7 (2024): pp. 3580–3599. DOI 10.1109/TKDE.2024.3352100.

[14] Sarica, Serhad, Luo, Jianxi and Wood, Kristin L. "TechNet: Technology Semantic Network Based on Patent Data." *Expert Systems with Applications* Vol. 142 (2020): p. 112995. DOI 10.1016/j.eswa.2019.112995.

[15] Fenoglio, Ennio, Kazim, Emre, Latapie, Hugo and Koshiyama, Adriano. "Tacit Knowledge Elicitation Process for Industry 4.0." *Discover Artificial Intelligence* Vol. 2 No. 1 (2022). DOI 10.1007/s44163-022-00020-w.

[16] Sanya, Richard and Shehab, Essam. "Towards a Knowledge Graph Framework for Ad Hoc Analysis in Manufacturing." *International Journal of Advanced Manufacturing Technology* (2023)DOI 10.1007/s10845-023-02319-6.

[17] Jia, Jiong, Zhang, Shuai and Saad, Mohamed. "Knowledge Graph-Based Multi-Granularity Tacit Design Knowledge Reuse." *Journal of Computational Design and Engineering* Vol. 12 No. 1 (2024). DOI 10.1093/jcde/qwae108.

[18] Kernan Freire, Samuel et al. "Tacit Knowledge Elicitation for Shop-Floor Workers with an Intelligent Assistant." *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023. DOI 10.1145/3544549.3585755.

[19] Goridkov, Nikita, Rao, Vivek, Cui, Di, Grandi, Daniele, Wang, Ye and Goucher-Lambert, Kosa. "Embedding Experiential Design Knowledge in Interactive Knowledge Graphs." *Journal of Mechanical Design* Vol. 145 No. 4 (2023): p. 041412. DOI 10.1115/1.4056347.

[20] Bolanos, Diego, Ataei, Milad, Grandi, Daniele and Goucher-Lambert, Kosa. "RECALL-MM: A Multimodal Dataset of Consumer Product Recalls for Risk Analysis Using Computational Methods and Large Language Models." *Proceedings of the ASME IDETC/CIE 2025*. DETC2025-169712. 2025.

[21] Zheng, Lianmin et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *Advances in Neural Information Processing Systems*. 2023. URL https://arxiv.org/abs/2306.05685.

[22] Szymanski, Mateusz et al. "Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks." *Proceedings of ACM IUI 2025*. 2025. DOI 10.1145/3708359.3712091.

[23] Chen, Guiming Hardy et al. "Humans or LLMs as the Judge? A Study on Judgement Biases." *arXiv preprint arXiv:2402.10669* (2024).

[24] Jiang, Weiwei et al. "A Survey on LLM-as-a-Judge." *arXiv preprint arXiv:2411.15594* (2024).

[25] Li, Dawei et al. "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods." *arXiv preprint arXiv:2412.05579* (2024).

[26] Edwards, Grace, Tehranchi, Fatemeh, Miller, Scarlett and Ahmed, Faez. "AI Judges in Design: Statistical Perspectives on Achieving Human Expert Equivalence with Vision-Language Models." *Proceedings of the ASME IDETC/CIE 2025*. 2025. URL https://arxiv.org/abs/2504.00938.

[27] Liu, Yang et al. "G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment." *Proceedings of EMNLP 2023*. 2023. URL https://arxiv.org/abs/2303.16634.

[28] Chong, Lena, Raina, Ayush, Goucher-Lambert, Kosa, Kotovsky, Kenneth and Cagan, Jonathan. "The Evolution and Impact of Human Confidence in Artificial Intelligence and in Themselves on AI-Assisted Decision-Making in Design." *Journal of Mechanical Design* Vol. 145 No. 3 (2023): p. 031401. DOI 10.1115/1.4055123.

[29] Zhang, Grace, Raina, Ayush, Cagan, Jonathan and McComb, Christopher. "A Cautionary Tale About the Impact of AI on Human Design Teams." *Design Studies* Vol. 72 (2021): p. 100990.

[30] Frich, Jonas et al. "AI-Augmented Brainwriting: Investigating the Use of LLMs in Group Ideation." *Proceedings of CHI 2024*. 2024.

[31] Kernan Freire, Samuel et al. "Knowledge Sharing in Manufacturing Using LLM-Powered Tools: User Study and Model Benchmarking." *Frontiers in Artificial Intelligence* (2024)DOI 10.3389/frai.2024.1293084.

[32] Tsakalerou, Mariza, Akhmadi, Saltanat, Balgynbayeva, Aruzhan and Kumisbek, Yerdaulet. "AI-Assisted Design Synthesis and Human Creativity in Engineering Education." *Frontiers in Artificial Intelligence* Vol. 9 (2026): p. 1714523. DOI 10.3389/frai.2026.1714523.

[33] Panahi, Sirous, Watson, Jason and Partridge, Helen. "Towards Tacit Knowledge Sharing Over Social Web Tools." *Journal of Knowledge Management* Vol. 17 No. 3 (2013): pp. 379–397. DOI 10.1108/JKM-11-2012-0364.

[34] Arab, Maryam, LaToza, Thomas D., Liang, Jenny and Ko, Amy J. "An Exploratory Study of Sharing Strategic Programming Knowledge." *Proceedings of CHI 2022*: pp. 66:1–66:15. 2022. DOI 10.1145/3491102.3502070.

[35] Fillmore, Charles J. "Frame Semantics." Linguistic Society of Korea (ed.). *Linguistics in the Morning Calm*. Hanshin Publishing, Seoul (1982): pp. 111–137.

[36] Hyland, Ken. *Hedging in Scientific Research Articles*. John Benjamins Publishing, Amsterdam (1998). DOI 10.1075/pbns.54.

[37] Toulmin, Stephen E. *The Uses of Argument*. Cambridge University Press, Cambridge (1958).

[38] OpenAI. "GPT-5 Model Card and System Documentation." OpenAI Platform Documentation (2025). URL https://platform.openai.com/docs/models/gpt-5.

[39] OpenAI. "New Embedding Models and API Updates: text-embedding-3-large." OpenAI Platform Documentation (2024). URL https://platform.openai.com/docs/guides/embeddings.

[40] Grootendorst, Maarten. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." *arXiv preprint arXiv:2203.05794* (2022).

[41] McInnes, Leland, Healy, John and Melville, James. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv preprint arXiv:1802.03426* (2018).

[42] McInnes, Leland, Healy, John and Astels, Steve. "hdbscan: Hierarchical Density Based Clustering." *Journal of Open Source Software* Vol. 2 No. 11 (2017): p. 205. DOI 10.21105/joss.00205.

[43] Gwet, Kilem Li. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." *British Journal of Mathematical and Statistical Psychology* Vol. 61 No. 1 (2008): pp. 29–48. DOI 10.1348/000711006X126600.

[44] Feinstein, Alvan R. and Cicchetti, Domenic V. "High Agreement but Low Kappa: I. The Problems of Two Paradoxes." *Journal of Clinical Epidemiology* Vol. 43 No. 6 (1990): pp. 543–549. DOI 10.1016/0895-4356(90)90158-L.

[45] Gao, Yanjun, Li, Ruizhe, Croxford, Emma, Caskey, John, Patterson, Brian W., Churpek, Matthew, Miller, Timothy, Dligach, Dmitriy and Afshar, Majid. "Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study." *JMIR AI* Vol. 4 (2025): p. e58670. DOI 10.2196/58670.

[46] Ma, Kevin, Grandi, Daniele, McComb, Christopher and Goucher-Lambert, Kosa. "Do Large Language Models Produce Diverse Design Concepts? A Comparative Study with Human-Crowdsourced Solutions." *Journal of Computing and Information Science in Engineering* Vol. 25 No. 2 (2025): p. 024501. DOI 10.1115/1.4067332.

[47] Doshi, Anil R. and Hauser, Oliver P. "Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content." *Science Advances* Vol. 10 No. 28 (2024): p. eadn5290. DOI 10.1126/sciadv.adn5290.

[48] Wang, Peng, Liu, Hongjun, Zou, Liye and Paas, Fred. "Overloaded Minds and Machines: A Cognitive Load Framework for Human-AI Symbiosis." *Artificial Intelligence Review* Vol. 59 (2026): p. 109. DOI 10.1007/s10462-026-11510-z.

[49] Song, Binyang, Soria Zurita, Nicolás, Gyory, Joseph et al. "When Faced with Increasing Complexity: The Effectiveness of AI Assistance for Drone Design." *Journal of Mechanical Design* (2022).

[50] Ariai, Alex et al. "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models and Challenges." *ACM Computing Surveys* (2025)DOI 10.1145/3716627.

[51] Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology*, 4th ed. SAGE Publications (2018).

[52] Fleiss, Joseph L. "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin* Vol. 76 No. 5 (1971): pp. 378–382. DOI 10.1037/h0031619.

APPENDIX A. PIPELINE DATA DIAGNOSTICS

This appendix summarizes the data volumes, distributions, and graph statistics produced by the DFM knowledge-extraction pipeline described in the main text. All counts reflect the final, merged extraction output after deduplication and quality filtering.

The pipeline begins by scraping question threads from five Stack Exchange sites via the public API. A GPT-based relevance filter retains only threads pertaining to design-for-manufacturability, reducing the corpus from 12,805 to 639 threads. Each retained thread is normalized to plain text, and both

TABLE 5: Pipeline stage summary from raw scrape to knowledge graph.

Stage	Count
Raw threads scraped	12,805
Posts with answers	1,399
Posts sent to LLM extraction	1,936
Unique threads filtered	911
Unique authors	1,138
Knowledge frames extracted	3,879
Avg. frames per post	3.22
Embedding vectors (3 collections)	11,637
Graph nodes	7,864
Structural edges	8,776
Semantic edges	4,372

questions and answers (1,936 posts total) are sent to a structured-output LLM call that extracts typed *knowledge frames*. The resulting frames, posts, authors, and threads form a provenance graph augmented with semantically inferred inter-frame edges.

A.1. GPT-5-mini Relevance Filter

A.1.1. Purpose. After keyword-based scraping returns 12,805 candidate threads, a GPT-5-mini call decides whether each thread is in scope for the DFM knowledge base. The filter is a binary classifier; it does not score or rank.

A.1.2. Model configuration. The filter uses the OpenAI Responses API with model `gpt-5-mini` at `reasoning_effort="low"` and `max_output_tokens=512`, with strict `json_schema` output enforcing the response shape `{"keep": bool, "reason": str}`. The reason field is constrained to ≤ 20 words. The filter runs with up to 10 parallel API workers in production and supports checkpoint resume to handle long runs.

A.1.3. Scope definition. The filter prompt is built from a fixed scope statement plus the candidate thread's metadata. The default scope statement used for the dataset reported in this paper is:

> *We want posts that provide practical knowledge, heuristics, risks, or principles about CNC machining and design for manufacturability (DFM) for machined parts. Include CAM/CAD/CAE and tooling details when they directly affect machinability. Include CNC coolant automation/workholding/fixturing topics when tied to machining practice. Include any CAD, CAM, or CAE topics (even if just software focused), and relevant mechanical design information. Exclude general electronics and unrelated mechanical topics.*

A.1.4. System prompt. The filter's system prompt instructs strict scope adherence and defaults to rejecting borderline cases:

> *You are a strict filter. Decide if a post is within scope. Return ONLY valid JSON:* `{"keep": true/false, "reason": "..."}`. *No code fences, no extra text. Reason must be ≤ 20 words. If unsure, set* `keep=false`.

TABLE 6: GPT-5-mini relevance-filter keep rates by Stack Exchange site, across 9,910 candidate threads with logged decisions. The sites with the highest keep rates (engineering, woodworking, robotics) align with the manufacturing-knowledge focus of the filter prompt.

Site	Candidates	Kept	Keep rate
engineering	839	452	53.9%
woodworking	493	35	7.1%
robotics	755	44	5.8%
electronics	7,396	102	1.4%
mechanics	427	6	1.4%
Total	**9,910**	**639**	**6.45%**

The bias toward rejection is intentional: the downstream extraction LLM call is expensive (and parses HTML, runs anti-hallucination quote verification, and applies seven additional structural filters), so the cost of a false-positive at the relevance stage is much higher than the cost of a false-negative.

A.1.5. Input fields. For each candidate thread, the filter receives the Stack Exchange `site` name, the question `title`, the question `tags`, and the first 8,000 characters of the question body (the body is truncated to keep prompt cost bounded; the truncated suffix rarely contains the load-bearing scoping signal). Answer bodies are not shown to the filter—answers are filtered indirectly through their parent question's keep decision.

A.1.6. Per-site keep rates. The filter's per-site behavior reflects the underlying content distribution on each Stack Exchange community. Table 6 reports keep rates across the 9,910 candidate threads with logged decisions.

The Engineering site has by far the highest keep rate (53.9%) because its tag taxonomy explicitly includes CNC machining, tolerancing, fixturing, and manufacturing-process tags; most of its candidate threads are squarely in scope. Electronics and Mechanics contribute the largest *candidate* pools but the smallest *kept* pools. Most electronics threads concern PCB design, power supplies, or signal integrity—adjacent to DFM but out of scope for this paper's machining focus. Most mechanics threads concern theoretical mechanics or vehicle dynamics rather than manufacturing practice.

A.1.7. Decision provenance. Every filter decision is logged with the model's `reason` string and persisted alongside the kept-thread list, supporting downstream auditing and any future re-scoping of the corpus without re-incurring the API cost. Representative reject reasons from the production run include: *"General electronics/power-supply question, not CNC machining/DFM content"*; *"Mechanism design question, not about CNC machining, DFM, or manufacturing"*; *"Mechanical mechanism question not related to CNC machining, DFM, CAM/CAD/CAE, or tooling."* Representative keep reasons include: *"Discusses tooling, speeds, and machinability for aluminum lathe work—practical machining/DFM topic"*; *"Discusses machining tolerances and design for manufacturability of shaft snap-ring grooves"*; *"Asks for machining physics, torque/power, and material effects—directly about machining and manufacturability."*

TABLE 7: Distribution of extracted knowledge frames by type.

Frame Type	Count	%
heuristic	1,401	36.1
observation	1,345	34.7
risk	428	11.0
principle	396	10.2
comparison	149	3.8
workaround	145	3.7
case	15	0.4

TABLE 8: Distribution of extracted frames by scope domain.

Scope Domain	Count	%
design_geometry	1,166	30.1
machining_process	581	15.0
other	512	13.2
machine_capability	507	13.1
quality	235	6.1
programming	225	5.8
tooling	225	5.8
material	165	4.3
shop_operations	141	3.6
cost	122	3.1

A.1.8. Cost and runtime. At GPT-5-mini's posted rates and the observed mean prompt size, the relevance filter cost $0.0005 per thread, or $5 to process the full 9,910-thread candidate pool. Wall-clock runtime with 10 parallel workers was 35 minutes. These costs are 1% of the downstream extraction cost, justifying the filter's role as a pre-extraction gate.

A.1.9. Limitations. The relevance filter has three known limits. First, the filter sees only the question body and tags; some threads with off-topic questions but on-topic answers will be rejected. Second, the 8,000-character body truncation can drop late-paragraph context on very long question posts. Third, the filter's bias toward rejection (instructed by the system prompt above) trades a higher false-negative rate for a lower false-positive rate, which suits the downstream cost asymmetry but limits absolute corpus completeness. We do not provide a ground-truth-validated precision/recall estimate of the filter in this paper; constructing one would require a hand-labelled gold set of ∼1,000 threads.

Each frame is classified into one of seven substantive types (an eighth, *other*, is reserved for edge cases and was not observed in the final dataset). **Heuristics** capture personal practices, rules of thumb, or directed advice together with a rationale. **Observations** record empirical, factual, or definitional claims that carry no explicit action or causal mechanism. **Risks** link a set of conditions to a failure mode or operational limitation. **Principles** provide causal explanations of *why* a phenomenon occurs. **Comparisons** weigh two or more alternatives against each other. **Workarounds** describe non-standard fixes for stated limitations. **Cases** are specific real-world job or incident narratives.

Scope domains indicate the primary technical area a frame addresses. The dominance of *design_geometry* (30.1%) reflects the pipeline's focus on DFM: many frames concern feature tolerances, wall thicknesses, hole placement, and other geometry-driven constraints. *Machining_process* and *machine_capability* together account for a further 28.1%, covering feeds, speeds, tool paths, and machine-specific limitations. The *other* category (13.2%) captures frames whose scope spans multiple domains or falls outside the predefined taxonomy.

TABLE 9: Epistemic stance distribution across extracted frames.

Epistemic Stance	Count	%
direct_neutral	2,619	67.5
tentative_subjective	1,022	26.4
hedged_emphatic	127	3.3
absolute_authoritative	94	2.4
narrative_implied	17	0.4

TABLE 10: Applicability levels assigned to extracted frames.

Applicability	Count	%
context_dependent	2,724	70.2
universal	1,115	28.7
situational	40	1.0

Epistemic stance captures the confidence and authority with which the original author presents a claim. The majority of frames (67.5%) are *direct_neutral*—stated matter-of-factly without hedging or strong emphasis. *Tentative_subjective* frames (26.4%) include qualifiers such as "I think," "usually," or "in my experience." *Hedged_emphatic* frames combine hedging language with strong conviction, while *absolute_authoritative* frames present claims as definitive rules. *Narrative_implied* captures knowledge embedded in storytelling rather than explicit assertion.

Applicability indicates how broadly a frame generalizes. *Context_dependent* frames (70.2%) apply only under specific conditions—such as a particular material, machine type, or feature geometry—and require an accompanying `example_context` field. *Universal* frames (28.7%) hold across a wide range of conditions without significant qualification. *Situational* frames (1.0%) are tied to a highly specific, one-off scenario.

The knowledge graph contains two categories of edges. **Structural edges** encode provenance: which post a frame was extracted from, which thread contains the post, who authored it, and the question–answer relationship between Stack Exchange posts. **Semantic edges** are inferred by a separate LLM pass that classifies the relationship between pairs of frames sharing similar subjects or embeddings. *SPECIFIES* indicates that one frame adds detail or constraints to another. *MITIGATES* links a frame (typically a heuristic or workaround) to a risk it addresses. *CAUSAL_LINK* connects frames sharing a cause–effect relationship. *CONTRADICTS* flags frames that present conflicting claims, enabling downstream systems to surface disagreements.

Posts were drawn from five Stack Exchange sites selected for their coverage of manufacturing, machining, and related engineering knowledge. The *Engineering* site contributes the ma-

TABLE 11: Edge types in the provenance knowledge graph.

Edge Type	Category	Count
POST_CONTAINS_FRAME	Structural	3,879
THREAD_CONTAINS_POST	Structural	1,936
POST_AUTHORED_BY	Structural	1,936
SE_ANSWERS	Structural	1,025
SPECIFIES	Semantic	2,159
MITIGATES	Semantic	1,218
CAUSAL_LINK	Semantic	637
CONTRADICTS	Semantic	358

TABLE 12: Distribution of posts sent to extraction by Stack Exchange site.

Stack Exchange Site	Posts	%
Engineering	1,345	69.5
Electronics	276	14.3
Robotics	174	9.0
Woodworking	106	5.5
Mechanics	35	1.8
Total	**1,936**	**100**

jority of DFM-relevant content (69.5%), while *Electronics* and *Robotics* provide complementary material on electromechanical integration and automation constraints. *Woodworking* contributes subtractive-manufacturing knowledge that transfers to CNC contexts, and *Mechanics* provides foundational physics and materials content.

The full knowledge graph is shown below:

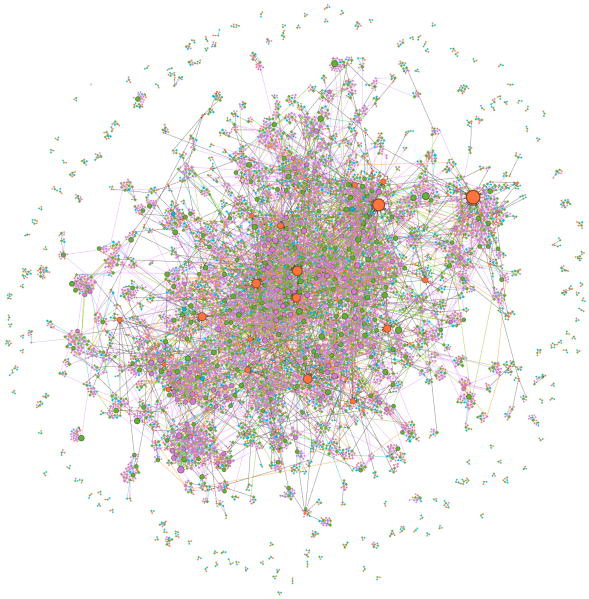

FIGURE 4: The full knowledge graph, with structural and semantic edges.

APPENDIX B. VALIDATION DETAILS: HUMAN PANEL AND LLM JUDGE

This appendix gives the per-rater, per-dimension breakdown of the human-expert frame review (Section B.1) and the full-corpus GPT-5 LLM-judge run that scored all 3,879 frames (Section B.2). The three-way pipeline ablation that isolates the contribution of retrieval, graph traversal, and structured synthesis is reported separately in Appendix C.

B.1. Human Review Panel

B.1.1. Panel composition (rotating design). Three engineering-reviewer slots (denoted R_a, R_b, R_c) rated a stratified sample of 27 knowledge frames drawn from 8 source posts. The panel is a *rotating design*: the original 26 frames were rated by one fixed triple of reviewers, and the supplemental Frame 27 was rated by a second triple drawn from a supplemental reviewer pool to close the post-level frame-type stratification gap. IRR is computed on the slot positions (R_a, R_b, R_c) under the rotating-panel assumption that the strictness profile of each slot is approximately constant across both panels; the recompute against the unified slot mapping produces IRR estimates that shift by at most ±0.02 AC1 relative to the original 26-frame panel. Two earlier reviewers were excluded prior to publication—one (R_0) was retained only for pilot calibration, and one ($R_{4'}$) produced ceiling-effect ratings across all dimensions and failed a held-out consistency check.

TABLE 13: Per-rater and panel-mean ratings (n=27 frames, 3 raters). EA / TK / CL are on a 1–3 scale; the remaining six are on a 1–5 scale.

Dimension	R_a	R_b	R_c	Mean	SD
Extraction accuracy	2.89	2.96	2.93	2.93	0.26
tacit knowledge	2.04	2.22	1.85	2.04	0.79
Clarity	2.89	2.93	2.93	2.91	0.32
Frame type	3.96	4.59	4.22	4.26	0.97
Post role	3.81	4.70	4.96	4.49	0.79
Scope	4.59	4.74	4.11	4.48	0.70
Applicability	4.52	4.67	3.81	4.33	0.86
Main point quality	4.78	4.89	4.74	4.80	0.43
Epistemic stance	4.52	4.89	4.22	4.54	0.93

The sampling design is post-based stratified by plurality `frame_type`. Forum-site stratification was not used; 2 of 5 sites (engineering, electronics) are represented in the panel. Closing the site-stratification gap would require an additional 5 frames (2 robotics, 2 woodworking, 1 mechanics) that are currently in supplemental review.

B.1.2. Rubric. Each frame was scored on nine dimensions. The first three (*extraction accuracy*, *tacit knowledge*, *clarity*) use a 1–3 scale capturing knowledge-content quality. The remaining six (*frame type*, *post role*, *scope*, *applicability*, *main point quality*, *epistemic stance*) use a 1–5 scale capturing label appropriateness against the KnowledgeFrame schema's controlled vocabularies. Each frame also received an overall *Keep* / *Revise* / *Remove* judgment. Full rubric text is in the supplemental data package.

B.1.3. Per-rater means. Table 13 reports per-rater and panel-mean ratings on all nine dimensions for the n=27 panel.

The three raters show stable cross-dimension profiles: R_a is the strictest on metadata classification (particularly post role);

TABLE 14: Overall verdict distribution per rater and by 3-rater majority (n=27 frames).

	Keep	Revise	Remove
R_a only	22 (81%)	5 (19%)	0
R_b only	23 (85%)	4 (15%)	0
R_c only	23 (85%)	4 (15%)	0
Majority	**25 (93%)**	**2 (7%)**	**0**

TABLE 15: Inter-rater reliability across the nine rubric dimensions (n=27 frames, 3 raters).

Dimension	Gwet AC1	Mean pairwise % agree
Extraction accuracy	0.85	90.1%
tacit knowledge	0.25	49.4%
Clarity	0.85	90.1%
Frame type	0.48	65.4%
Post role	0.72	81.5%
Scope	0.81	87.7%
Applicability	0.56	70.4%
Main point quality	0.96	97.5%
Epistemic stance	0.68	79.0%

R_b is the most generous across the metadata dimensions; R_c is the most critical on tacit-knowledge value (mean 1.85/3) and applicability (mean 3.81/5). No rater systematically dominates ratings on the content-quality dimensions, which cluster near the ceiling for all three raters.

B.1.4. Overall verdicts. Each rater's individual verdict distribution and the 3-rater majority verdict are given in Table 14. No frame was flagged *Remove* by any rater.

The two majority-Revise frames were F5 and F6, both flagged by at least two raters for possible main-point hallucination (the system-generated summary included detail the source quote did not clearly support). The supplemental Frame 27 received a unanimous *Keep* from all three raters.

B.1.5. Inter-rater reliability. Inter-rater agreement was assessed using Gwet's AC1 [43] on tier-collapsed *Poor* (1–2) / *Acceptable* (3) / *Good* (4–5) categories, with the chance-agreement term computed using a fixed $K = 3$ to avoid undefined values on dimensions with concentrated ratings. Pairwise percent agreement on the same tier-collapsed scale is reported alongside for interpretability. Table 15 summarizes both metrics.

We adopt AC1 rather than Krippendorff's α [51] or Fleiss' κ [52] as our primary IRR statistic. Both α and κ are well-known to produce paradoxically low values when one category dominates the marginal distribution—a property known as the "kappa paradox" [44]. In our panel, 88–99% of ratings on the content-quality dimensions fall in the *Good* tier, which mathematically suppresses α and κ toward zero even at near-perfect raw agreement. For example, the 96% all-three-rater exact agreement on `main_point_quality` produces a Fleiss κ in the slightly-negative range, which mis-states the panel's actual consensus on the dimension. AC1 corrects for this by computing the chance-agreement baseline from the marginal proportions themselves rather than assuming uniform-marginal chance agreement, which is the source of the paradox.

We do report Krippendorff's α (ordinal weighting) in the supplemental data package for completeness, since it is the standard statistic for ordinal-rating IRR in much of the social sciences and clinical-NLP literature. Doug McGowan computed α for the original 26-frame panel; the values are reported in the corpus diagnostic logs at `dfm_scraping/IRR_Metrics.md`. The pattern is consistent with AC1 in ranking dimensions but compressed toward zero by a factor of 0.5–0.8 due to the ceiling-distribution effect described above. We therefore do not use α as the primary statistic in the main text or this appendix; it appears only as a robustness-check entry in the supplemental table. The same caveat applies to Fleiss κ, which shows the same pattern at slightly more extreme compression because κ uses a fixed K = observed-categories denominator that further penalizes the ceiling case.

The pattern in Table 15 is methodologically informative. Extraction accuracy, clarity, and main point quality have AC1 ≥ 0.85, indicating strong consensus that the extraction pipeline produces faithful, well-written knowledge claims. Classification labels are noisier: frame type (0.48) and applicability (0.56) fall below the conventional 0.65 threshold for authoritative use. We therefore present `frame_type` and `applicability` as advisory metadata rather than authoritative filters in the main text (Section 5.3). tacit knowledge (0.25) is the noisiest dimension; the *textbook* versus *tacit* boundary is inherently subjective and domain-dependent.

TABLE 16: LLM-judge verdict distribution across the full 3,879-frame corpus. GPT-5 default reasoning effort, identical rubric to the human panel.

Verdict	Count	%
Keep	3,103	80.0
Revise	771	19.9
Remove	5	0.1

B.2. GPT-5 LLM-Judge: Full-Corpus Review

B.2.1. Setup. To scale frame-level quality review beyond the 27-frame human panel, we used GPT-5 (default reasoning effort) as a calibrated LLM judge [21] on all 3,879 extracted frames. The judge applies the same rubric as the human panel (Section B.1), with structured Pydantic output enforcing the rating ranges and verdict categories. Each judge call receives (i) the source post text excluding any block quotes, (ii) the verbatim `source_quote` extracted by the pipeline, and (iii) the full structured frame including all classification labels.

B.2.2. Verdict distribution. Table 16 reports the verdict distribution across the full corpus. The LLM judge flags roughly one frame in five for revision, and effectively zero frames for removal—consistent with the human panel's view that the pipeline does not produce frames that warrant complete rejection.

B.2.3. Per-dimension descriptive statistics. Table 17 reports mean, median, and standard deviation for the nine rubric dimensions across the 3,879 LLM-judge ratings.

The rating-bin distributions reveal where the LLM judge agrees and disagrees with the human panel. On extraction accuracy, 92.4% (3,586/3,879) of frames receive the top rating, match-

TABLE 17: LLM-judge per-dimension descriptive statistics (n=3,879 frames). EA / TK / CL on the 1–3 scale; remaining dimensions on the 1–5 scale.

Dimension	Mean	Median	SD
Extraction accuracy	2.92	3	0.27
tacit knowledge	1.71	2	0.56
Clarity	2.99	3	0.08
Frame type	4.75	5	0.65
Post role	4.94	5	0.31
Scope	4.33	5	0.89
Applicability	4.51	5	0.88
Main point quality	4.58	5	0.55
Epistemic stance	4.65	5	0.86

ing the human ceiling effect. On clarity, 99.5% (3,858/3,879) receive the top rating, again matching humans. Classification labels are more dispersed: frame type rates 5 on 82.7% of frames but assigns 4 on 12.7% and ≤ 3 on 4.6%; applicability rates 5 on 69.9% but spreads the remaining 30.1% across all four lower categories. Scope shows the widest dispersion among the metadata dimensions (SD 0.89), with 47 frames rated as a 1.

B.2.4. Comparison to the human panel. On the 27-frame overlap, the LLM judge is systematically harsher than the human panel on overall verdicts (LLM: 46% Keep on the 26-frame original sample; human majority: 88% Keep), driven primarily by stricter assessment of `scope` ($\Delta_{\text{LLM-human}} = -0.75$ on the 1–5 scale). Near-zero mean differences appear on clarity (0.03) and post role (0.02), indicating that the judge agrees with humans on the load-bearing content-quality and post-classification questions. Section 5.2 of the main text discusses this asymmetry and its implication for using the LLM-judge verdict as an automated filter.

B.2.5. Intra-judge repeatability (n=10 frames × 3 runs). To isolate intra-model stochasticity, GPT-5 (reasoning effort low, JSON schema output) was queried three times per frame on a random 10-frame subsample under identical prompts and varying random seed. Results in Table 18. All five classification-label dimensions and clarity show 100% identical ratings across all three runs. The overall Keep/Revise/Remove verdict was identical across all three runs on 5 of 10 frames (50%), with the remaining 5 frames showing a 2-of-3 majority verdict driven by single-run shifts in content-quality dimensions. We therefore present the full-corpus LLM-judge verdicts (Table 16) as single-run statistics; for downstream applications requiring deterministic filtering, a majority-of-N voting scheme would be the standard mitigation.

B.2.6. Judge-model ablation (n=1,150 partial). A judge-model ablation was run on 1,150 of the 3,879 frames using GPT-5.4 at medium reasoning effort, under identical prompt, schema, and seed. Cross-judge agreement is strong on content-quality dimensions (mean absolute difference of 0.01 on clarity and 0.19 on extraction accuracy, both on the 1–3 scale) and weaker on classification-label dimensions (0.61 on applicability, 0.51 on main-point quality, 0.46 on frame type, all on the 1–5 scale). The dimensions where the two LLM judges disagree most are the same dimensions where the human panel disagrees most, providing convergent evidence that classification labels are rubric-underdetermined for both human and automated raters. The ablation data and a full per-dimension comparison are in the supplemental data package.

TABLE 18: Intra-judge repeatability across 3 GPT-5 runs on 10 frames under identical prompt and varying random seed. "Identical across 3" = fraction of frames where all 3 runs returned the same rating.

Dimension	Identical across 3
extraction_accuracy	60%
tacit_knowledge	70%
clarity	100%
frame_type_rating	100%
post_role_rating	100%
scope_rating	100%
applicability_rating	100%
main_point_quality	50%
epistemic_stance_rating	100%
Overall verdict	**50%**

B.2.7. Known limits. GPT-5 may exhibit self-enhancement bias [21] when judging outputs from a model in the same family that performed the extraction (gpt-5-nano in our pipeline; see Section A.1). The judge-model ablation above provides a partial test: if self-enhancement dominated, gpt-5.4 should diverge most on extraction accuracy, where the judge's own family produced the frame. Instead, divergence concentrates on classification labels, which are rubric-driven rather than extraction-driven. Judge ratings reflect rubric structure, not model-family agreement.

APPENDIX C. PIPELINE ABLATION: LLM-ONLY, LLM+RAG, FULL PIPELINE

This appendix reports the verbosity-controlled pipeline ablation used to isolate the contribution of vector retrieval, semantic-graph traversal, and structured synthesis. Eight queries spanning three corpus-coverage tiers were run against three configurations under a common JSON output schema and a common reasoning budget; per-query results are in Table 21 and per-config averages in Table 20.

C.1. Configurations and Controls

C.1.1. Backbone. All three configurations use GPT-5 at `reasoning_effort="low"` and a fixed `max_completion_tokens=4096` cap covering both hidden reasoning tokens and visible JSON output. Each configuration emits the same response schema: a list of design rules (up to ten), a list of knowledge gaps (up to ten), a list of cited frame IDs (up to twelve), and a summary capped at 200 words. The shared schema controls for the output-length confound that biases "richer answer" metrics in unconstrained baselines.

C.1.2. Question set. Eight queries were drawn across three coverage tiers. The *well-covered* tier (Q1 thin-walled aluminum, Q2 climb vs. conventional milling, Q3 deep pockets and tall narrow ribs) targets topics with dense corpus support. The *sparse* tier (Q4 chamfer cost impact, Q5 shop-floor traceability, Q6 tolerance stack-up) targets topics with only adjacent corpus material. The

TABLE 19: Ablation configurations. All three use GPT-5 at `reasoning_effort="low"` and emit the same JSON schema. Configuration C delegates to the full agent pipeline described in Section 4.3.

Config	Retrieval	Synthesis context
A. LLM-only	none	no context
B. LLM+RAG	top-12 vector hits	retrieved-frame summaries
C. Full pipeline	multi-vector + 2-hop graph	structured DFM report (resynth)

out-of-scope tier (Q7 sheet-metal air bending, Q8 FDM overhang support) is composed of designed-to-fail probes: neither sheet-metal forming nor additive manufacturing is represented in the corpus.

C.1.3. Configuration C resynth step. Configuration C's native output is a verbose structured Pydantic report (sub-aspects, evidence claims, reasoning chains, formulas, contradictions, gaps). To keep the visible output band comparable to A and B, a second LLM call compresses C's report into the common schema after the pipeline runs. C's *total* token spend (retrieval prompts, decomposition, synthesis, resynth) remains substantially higher than A or B; the resynth step matches only the visible-output verbosity.

C.2. Per-Configuration Behavior

Table 20 reports per-configuration averages across the eight queries. The three configurations occupy clearly distinct operating regions.

C.2.1. Configuration A behavior. Configuration A produces exactly ten design rules and (with one exception on Q7) ten knowledge gaps on every query, regardless of coverage tier. No retrieved frame IDs are cited because no retrieval is performed. The rules are drawn entirely from gpt-5's parametric knowledge and bear no relation to the corpus content, which is by design for the baseline. The constant 10/10 output shape reflects the schema constraint rather than calibrated evidence: A populates ten gaps even on a well-covered question because the schema requires the field.

C.2.2. Configuration B behavior. Configuration B retrieves twelve top-similarity frames per query and exposes their headline metadata to the LLM (frame ID, subject, frame type, scope, main point). The system prompt instructs the model to use only the retrieved-frame content and to cite each claim with a frame ID. Under those constraints the LLM judges the exposed payload too thin to compose design rules on six of the eight queries (Q1, Q3, Q5–Q8) and routes everything to the `knowledge_gaps` channel. On Q2 and Q4 it produces 7 and 0 rules, respectively, while citing 12 retrieved frames. This behavior is an implementation-level constraint of the v2 RAG baseline rather than a property of vector retrieval in general: exposing the full frame body (source quote, example value, example context, epistemic stance) and relaxing the strict citation requirement would likely shift B closer to the parametric-LLM end of the spectrum. We retain the strict-RAG variant here because it provides a clean lower bound on what retrieval-grounded answer composition can do when the retrieval payload contains only headline metadata.

C.2.3. Configuration C behavior. Configuration C cites between six and twelve retrieved frame IDs on every query. Its design-rule output ranges from five to ten per query (mean 8.25), with the lower counts concentrated on the out-of-scope queries (Q7 = 7 rules, Q8 = 5). The `knowledge_gaps` field ranges from five to seven entries per query, with the highest counts on the out-of-scope tier. Unlike A and B, C's gaps are domain-specific rather than schema-padded: on the FDM overhang query (Q8), gaps include "PLA-specific self-support angle thresholds for overhangs greater than 45 degrees and how they vary with temperature and cooling" and "Data quantifying nozzle diameter and layer height effects on maximum printable overhang angle or bridge length in PLA." On the sheet-metal query (Q7), gaps name "No numeric minimum bend radius values by stainless grade/temper (e.g., 304/316 annealed vs. half-hard) for 16-gauge" and "No guidance on grain direction effects (along vs. across rolling direction) on cracking risk in stainless bends."

C.3. Per-Query Results

Table 21 reports per-query latency, visible-token count, cited-frame count, rule count, and gap count for all 24 runs. Note the regularity of A's 10/10 output shape, B's collapse to zero rules on six of eight queries, and C's stable ≥ 6 cited frames on every query.

C.4. Out-of-Scope Behavior: A vs. C

The two out-of-scope queries (sheet-metal air bending; FDM overhangs greater than 45° in PLA) provide the cleanest discrimination between configurations because the corpus contains no frames on either subject.

Configuration A produces ten confident, well-formed answers on each query and lists ten generic "knowledge gaps" that read as topic extensions a user might want to know. The rules include specific numeric prescriptions drawn from gpt-5's parametric knowledge ("For 16-gauge stainless steel (~1.5 mm, 304/316), in air bending the minimum bend radius typically lies between 1.0× and 1.5× the material thickness"); the gaps describe classes of follow-up questions rather than corpus limitations. The output offers no signal to the user that the corpus is silent on either subject.

Configuration C produces fewer rules (seven on Q7, five on Q8) and the rules it does produce are framed in general engineering terms transferable from analogous corpus material rather than as domain-specific numeric prescriptions. Q7's rules emphasize strain-driven minimum-radius reasoning and overbend-for-springback control rather than chart values for 304/316. Q8's rules emphasize deflection-limited overhang design and unsupported-span control rather than PLA-specific 45° thresholds. Critically, C's gap list names *what is missing in this corpus*, not what is absent from general knowledge: "No numeric minimum bend radius values by stainless grade/temper for 16-gauge," "No guidance on grain direction effects on cracking risk in stainless bends," "PLA-specific self-support angle thresholds for overhangs greater than 45 degrees," "Quantitative guidance on cooling airflow placement and intensity needed to solidify PLA overhangs before sag occurs." This pattern is consistent

TABLE 20: Per-configuration averages across eight queries. "Visible tokens" is the completion-token count returned to the response schema. "Total tokens" includes hidden reasoning and pipeline overhead. "Cited > 0" counts queries in which the configuration cited at least one retrieved frame ID.

Config	Latency (s)	Visible tokens	Total tokens	Cited > 0	Mean rules	Mean gaps
A. LLM-only	83	1,212	1,212	0 / 8	10.00	9.75
B. LLM+RAG	156	1,125	1,125	2 / 8	0.88	10.00
C. Full pipeline	207	2,049	17,207	8 / 8	8.25	5.88

TABLE 21: Per-query ablation results (24 runs total: 8 queries × 3 configurations). "Cited" = number of retrieved frame IDs included in the cited_frame_ids field. "Rules" and "Gaps" are counts of returned design-rule and knowledge-gap entries respectively. Latency for A and B is the single API call; for C it is the full pipeline including retrieval, graph traversal, synthesis, and resynth.

Query	Coverage	Config	Latency (s)	Cited	Rules	Gaps
Q1 thin-walled aluminum	well-covered	A	60	0	10	10
	well-covered	B	175	0	0	10
	well-covered	C	201	12	10	6
Q2 climb vs. conventional	well-covered	A	53	0	10	10
	well-covered	B	184	12	7	10
	well-covered	C	227	9	9	5
Q3 deep pockets, narrow ribs	well-covered	A	85	0	10	10
	well-covered	B	144	0	0	10
	well-covered	C	211	11	9	6
Q4 chamfer cost impact	sparse	A	109	0	10	10
	sparse	B	156	12	0	10
	sparse	C	173	12	8	6
Q5 shop-floor traceability	sparse	A	97	0	10	10
	sparse	B	167	0	0	10
	sparse	C	220	10	10	6
Q6 tolerance stack-up case	sparse	A	79	0	10	10
	sparse	B	155	0	0	10
	sparse	C	238	10	8	5
Q7 sheet-metal bend (OOS)	out-of-scope	A	121	0	10	8
	out-of-scope	B	130	0	0	10
	out-of-scope	C	169	7	7	6
Q8 FDM overhang (OOS)	out-of-scope	A	60	0	10	10
	out-of-scope	B	138	0	0	10
	out-of-scope	C	220	6	5	7

with the design intent of the structured `knowledge_gaps` channel: when the corpus lacks evidence for a specific dimension of the query, the agent flags the missing dimension explicitly rather than masking the omission with confidence.

This distinction matters for downstream advisory use. Configuration A provides parametric advice indistinguishable from a base LLM; the user must know to verify against external sources. Configuration C provides corpus-grounded advice paired with an explicit audit trail of corpus limitations; the user is told both what is supported and what is not.

C.5. Cost, Latency, and Caveats

C.5.1. Cost and latency. Mean per-query cost is approximately $0.012 for Configuration A, $0.011 for Configuration B, and $0.15 for Configuration C; mean wall-clock latency is 83 s, 156 s, and 207 s respectively. C's 14× higher token spend and 2–3× higher latency relative to A and B are driven by graph traversal, query decomposition, structured synthesis, and the resynth step. Embedded cost numbers exclude figure and embedding regeneration costs paid once at corpus build time.

C.5.2. Refusal detection. The current implementation uses a text-pattern matcher (searching for phrases such as "no direct corpus evidence," "corpus does not," "not in the corpus") on the `knowledge_gaps` field to flag refusals. This matcher returns zero refusals for every record in the v2 run. Manual inspection of C's Q7 and Q8 outputs shows that C *does* implicitly refuse by listing corpus-specific missing facts, but in phrasing the keyword matcher does not catch (e.g., "No numeric minimum bend radius values by stainless grade/temper for 16-gauge" instead of "no direct corpus evidence on sheet-metal bending"). A more rigorous evaluation would parse the structured `knowledge_gaps` field with a separate classifier or have a human rater label coverage-correctness on each gap entry.

C.5.3. Configuration B's restrictive payload. Configuration B exposes only frame headline metadata (`frame_id`, `subject`, `frame_type`, `scope`, `main_point`) to the LLM, not the full frame body. The system prompt requires that every claim be cited to a frame ID and that no information outside the retrieved frames be used. Under those joint constraints the LLM produces zero design rules on six of the eight queries. Exposing

additional frame fields (`source_quote`, `example_value`, `example_context`, `epistemic_stance`) and relaxing the strict-citation requirement would likely produce a less conservative B that interpolates more freely from the retrieved context. We report the strict-RAG variant here as a lower bound on how a retrieval-only system behaves when the LLM is given only frame headlines.

C.5.4. Sample-size and model-coverage limits. The eight-query design probes operating regions of A, B, and C rather than producing precise mean estimates. Per-query numbers should be read as illustrative. All three configurations use the same backbone model (GPT-5); the differences are in retrieval and synthesis structure, not in model capability. An ablation that swaps backbone models is reported in Section B.2 for the judge role; an analogous backbone ablation for the synthesis role is left to future work.

APPENDIX D. EXAMPLE AGENT OUTPUT: CLIMB VS. CONVENTIONAL MILLING

This appendix presents a DFM agent response to the query *"How do I choose between climb and conventional milling?"*, generated using GPT-5 at default reasoning effort. Each section is annotated with its data source:

- **[KG]** Content traced directly to extracted knowledge frames in the graph.
- **[LLM]** Content generated by the language model from general engineering knowledge.
- **[KG+LLM]** LLM synthesis constrained to cite specific frames.

D.1. Summary

[KG+LLM] The summary is generated after all evidence has been retrieved, traversed, and ranked. It synthesizes across all disciplines into a concise answer.

> Climb milling favors accuracy and finish on rigid, backlash-free machines; conventional milling is safer on flexible or backlash-prone setups. Across both, ensure adequate chip load to avoid rubbing/heat, provide chip evacuation, and anticipate finish/burr implications when setting tolerances and edge-breaks.

D.2. Discipline Sections

The agent organizes evidence by engineering domain, selecting disciplines from the `ScopeDomain` vocabulary in the KnowledgeFrame schema. For this query, five disciplines were populated: *machining_process*, *machine_capability*, *tooling*, *quality*, and *design_geometry*. One is shown below in full.

D.2.1. Machining Process

First-Principles Rationale. **[LLM]** This field is generated entirely from the LLM's general engineering knowledge, not from frames. The prompt requires it to be clearly labeled as inference. For this discipline, the agent produced:

> Milling direction changes how the cutter's tangential force vector loads the work and machine drive. On flexible setups, climb's pulling force can draw the tool into the cut and amplify trajectory error, while conventional tends to push back against the drive. Too small a chip load leads to rubbing rather than shearing, raising temperature and provoking work hardening and chatter.

Governing Formulas. **[LLM]** Equations the LLM deems relevant to the discipline, not sourced from frames. Three were produced for this discipline, including cantilever deflection ($\delta = FL^3/3EI$), total compliance ($\delta_{\text{total}} = \delta_{\text{tool}} + \delta_{\text{work}}$), and a natural frequency scaling rule.

Evidence. **[KG]** Each claim is drawn directly from a knowledge frame's `main_point` field. The agent retrieved 8 evidence claims for this discipline; one is shown with its source frame data.

Claim	On machines with flexible gantries or slides, climb milling's force can pull the cutter off the programmed path.
Frame ID	`se-electronics-364080-a364100-2`
Type / Scope	`principle` / `machine_capability`
Applicability	`context_dependent`
Source Quote	"Or you may be doing climb milling and the cutter runs well outside of the desired path if the gantry and slides are too flexible."
Author	Spehro Pefhany (455K rep, high credibility)

Reasoning Chains. **[KG]** The agent traverses semantic edges in the knowledge graph to construct causal sequences. Two chains were produced for this discipline; one is shown below.

Too-light cuts lead to burnishing instead of cutting	
Root cause	Cuts that are too light allow elastic deflection so the edge does not bite properly. Frame `se-engineering-55277-a55280-0` (observation, DKNguyen, medium credibility)
CAUSAL_LINK →	
Effect	When depth becomes too shallow, the tool stops cutting and instead burnishes the surface. Frame `se-engineering-2413-a40671-4` (observation, Pekka, low credibility)

The CAUSAL_LINK edge connecting these frames was classified during the semantic edge inference stage of the pipeline, not by the agent at query time. The agent traverses pre-existing edges to assemble chains.

D.3. Design Rules

[KG+LLM] Design rules are actionable recommendations that the LLM synthesizes from evidence frames. Unlike evidence claims, which describe what the knowledge graph says, design rules state what the *designer* should do about it—the prompt constrains them to decisions made at the CAD or drawing stage. Eight rules were produced for this query; one is shown with its source frame.

Design Rule	When designing parts for flexible gantry/bench-top machines, **add local stiffness or machining stock** near critical contours because climb milling forces can pull the tool off-path on compliant structures.
Source Frame	`se-electronics-364080-a364100-2`
Frame Type	`principle`
Source Quote	"Or you may be doing climb milling and the cutter runs well outside of the desired path if the gantry and slides are too flexible."
Main Point	During climb milling on a machine with a flexible gantry or slides, cutting forces can pull the tool outside the programmed path. Toolpath fidelity depends on machine stiffness.

Note that the evidence claim and design rule above cite the same frame. The evidence claim restates the frame's observation; the design rule transforms it into a geometric action the designer controls.

D.4. Knowledge Gaps

[LLM] The agent identifies areas where the knowledge graph lacks sufficient evidence to fully answer the query. These gaps serve as both a transparency mechanism for the user and as targets for future data collection.

1. No direct comparative evidence on chip-thickness evolution for climb vs. conventional milling and its effect on heat.

2. No explicit evidence on built-up edge (BUE) tendencies vs. milling direction for different materials.

3. No quantified rigidity or backlash thresholds at which to prefer climb over conventional; guidance is qualitative only.

4. No material-specific direction preference (e.g., aluminum vs. steel) supported by the extracted frames.

5. No direct evidence on burr formation differences between climb and conventional side milling.