NIST Special Publication 1500
NIST SP 1500-37

Advancing Technical Language Processing and Large Language Models in Industrial Applications

Insights from TLP COI Events in 2024

Michael K. Dawson Jr.
Sarah Lukens
Michael E. Sharp

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-37

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication 1500
NIST SP 1500-37

Advancing Technical Language Processing and Large Language Models in Industrial Applications

Insights from TLP COI Events in 2024

Michael K. Dawson Jr. [1], Sarah Lukens [2], Michael E. Sharp [1]

[1] Smart Connected Systems Division, Communications Technology Laboratory, NIST, Gaithersburg, Maryland, USA
[2] LMI, Tysons, Virginia, USA

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-37

December 2025

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST Technical Series Publications: https://www.nist.gov/nist-research-library/nist-publications

SP 1500 Series

Publications in the SP 1500 subseries are intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the United States Government.

Publication History

Approved by the NIST Editorial Review Board on 2025-08-22.

Suggested Citation

Michael K. Dawson Jr., Sarah Lukens, Michael E. Sharp (2025). Advancing Technical Language Processing and Large Language Models in Industrial Applications: Insights from TLP COI Events in 2024. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Special Publication 1500 (SP)1500-37. https://doi.org/10.6028/NIST.SP.1500-37

Author Names and ORCID Identifiers

Michael K. Dawson Jr. (0000-0002-7151-5664); Sarah Lukens (0000-0003-1514-8721); Michael E. Sharp (0000-0002-8014-3612).

Contact Information: michael.dawson@nist.gov

Abstract

It remains challenging to integrate artificial intelligence (AI) and natural language processing (NLP) into complex engineering systems. Popular models and techniques are designed for use with everyday, non-technical text, and therefore perform poorly when attempting to interpret language containing specialized terminology or context-specific meanings. Technical Language Processing (TLP) attempts to address these challenges by tailoring NLP approaches for use in technical and engineering domains. As a nascent field, community-driven events have played a crucial role in facilitating dialogue on emerging issues and potential solutions.

In 2024, the TLP Community of Interest (COI), overseen by the National Institute of Standards and Technology (NIST), organized two events: a two day virtual workshop hosted by NIST, and a panel at the Annual Conference of the Prognostics and Health Management (PHM) Society. These two events provided opportunities for experts from academia, industry, and government to discuss the application of TLP and Large Language Models (LLMs) in industrial settings. This paper summarizes the key discussions and insights from the events, while highlighting the community's collective efforts to advance the application of TLP and LLMs in the industrial sector.

Keywords

Artificial intelligence; Data fusion; Engineering design; Industry; Information retrieval; Knowledge extraction; Large language models; Maintenance; Manufacturing; Natural language processing; Prognostics and health management; Retrieval-augmented generation; Standards; Technical language processing; Vision language models.

Table of Contents

List of Tables

Acknowledgments

The authors extend their appreciation to the many speakers and panelists who generously shared their insights and knowledge at our events. Similarly, we gratefully acknowledge the events' attendees, for their enthusiasm and participation throughout each event. We extend our gratitude to those who made these events possible, especially: the PHM Society, NIST Conference Services, David Wollman, and our fellow organizers. Finally, we thank our reviewers, Mehdi Dadfarnia, Thomas Roth, and Rachael Sexton, for their feedback on this report.

1. Introduction

As recent advancements drive the growth of industrial systems in size and complexity, the need for effective approaches to analyze vast amounts of technical text has become critical. Industries possess a vast amount of technical language data that range from detailed technical manuals to maintenance logs, all of which contain valuable information that can enhance operational efficiency and decision-making processes. Technical Language Processing (TLP) is an approach that adapts natural language processing (NLP) methodologies to the specific needs of technical and engineering domains [1, 2].

The assumptions underpinning NLP methods, while effective for general-purpose language tasks, can lead to poor analysis of text from technical domains. TLP is intended to, in part, formalize the process of examining these assumptions to ensure relevant interpretations of industry-specific terminology and documentation. But doing so takes time, and given the rapid pace of NLP development, including and especially the introduction of Large Language Models (LLMs), adapting state-of-the-art methods to industrial requirements in a timely manner is challenging. LLMs, in particular, offer enhanced capabilities for language understanding and generation, yet most “off-the-shelf” models lack domain-specific knowledge, are vulnerable to hallucinations, and unable to verify facts—limitations which, until addressed through TLP, make their use in high-stakes industrial settings ill-advised [3].

To advance research and practice in this area, the Technical Language Processing Community of Interest (TLP COI) brings together professionals from industry, academia and government. Past community events have shown to be informative, such as a 2020 workshop hosted at the National Institute of Standards and Technology (NIST) on gathering standards requirements for natural language analysis in manufacturing [4].

In the fall of 2024, the TLP COI organized two events to share emerging tools, applications and challenges in this field. The first was a two-day virtual workshop hosted by NIST in September [5]. The second was a panel held at the annual conference of the Prognostics and Health Management (PHM) Society in November, which discussed TLP and LLMs [6]. This paper presents insights and trends from both events.

The rest of this paper is organized as follows. A background of TLP is covered in Section 2. A summary of the TLP COI Workshop is covered in Section 3, which includes a detailed summary of each of the presentations and discussions. A summary of the panel at the PHM Society Conference with discussions and main points covered is presented in Section 4. A summary of the key takeaways from the events is in Section 5, and Section 6 concludes the report.

2. Background

2.1. Characteristics of TLP

One of the primary distinguishing characteristics of TLP is the emphasis on incorporating domain expertise into the language processing pipeline. Unlike generic NLP, which often relies on broad, general-purpose models, TLP integrates technical knowledge and terminology specific to a given domain. This is achieved through the development and utilization of technical dictionaries, customized tokenization strategies, and domain-specific embeddings that capture the unique linguistic features and nuances of technical texts [7–11].

In many cases, TLP applications lack large quantities of relevant data needed for standard NLP tool development [12–14]. This necessitates clever investment in and use of resources both to train and deploy TLP models and services. The development of TLP resources is inherently iterative and collaborative, involving close interaction between practitioners, domain experts, and stakeholders to ensure that the resulting models and tools accurately reflect the needs and characteristics of the target domain, which can vary significantly across different industries and applications.

A key aspect of TLP is its emphasis on maintaining a human-in-the-loop process for verifying the accuracy and trustworthiness of analyses. This involves ongoing validation and refinement of TLP models and outputs by domain experts, ensuring that the insights generated are not only technically accurate but also relevant and actionable within the specific context. The human-in-the-loop approach also facilitates the identification and mitigation of potential biases and errors, further enhancing the reliability of TLP-based analyses [15–17].

2.2. Applications of TLP

TLP supports a wide range of applications by combining domain expertise with computational methods. Here applications are summarized in three areas: TLP for data quality improvements, TLP for prescriptive and recommendation systems, and human-machine communication.

TLP enables the extraction of actionable insights from unstructured and semi-structured data sources, such as maintenance logs, inspection reports, and operational records. This can assist management of large, unstructured and potentially dirty data through summarization or by extracting structured fields from the unstructured text [13, 18, 19]. Failure Mode Classification is a specific task which labels unstructured work orders with failure modes [14, 20]. Structured maintenance data can be used to estimate input parameters for reliability metrics or statistical analyses [8, 10, 21] or for Failure Modes and Effects Analysis (FMEA) [22]. Approaches which focus on bridging the gap between engineering FMEA templates and maintenance data for providing quantitative risk estimates [23, 24].

TLP also supports prescriptive tasks, acting as recommendation systems for troubleshooting, diagnostics and defect inspection within PHM [9, 25–29]. Recommendation systems in adjacent application include real time suggestion systems which utilize structured databases of historical failures and actions taken to make recommendations for maintenance actions [30, 31] and for manufacturing operations [32].

Root cause analysis is an application where TLP supports the development and access to a knowledge framework in manufacturing [33, 34]. These frameworks can be developed and utilized for design knowledge retrieval [35], assessing production issues [36], and for identifying and reducing maintenance and defect-related occurrences [37, 38].

TLP enhances communication between humans and machines through the understanding of technical documents, structured procedures, engineering specifications, and other domain-specific materials [39–42]. Developments of note include creation of ontologies specific to maintenance data and fault diagnosis [22, 39, 43], as well as maintenance actions [40] and maintenance procedures [44].

2.3. Advancements in NLP and the Rise of the LLMs

Recent advancements have focused on adapting LLMs to technical application through techniques such as fine-tuning, domain-specific data augmentation, and Retrieval-Augmented Generation (RAG), where the LLM is augmented with external knowledge sources to improve the accuracy and relevance of generated text [45, 46]. Emerging approaches like GraphRAG further enhance LLMs by integrating graph-structured data for better reasoning over complex technical relationships [47]. Additionally, multimodal models that combine text with other data types, such as images or sensor streams, are expanding the reach of TLP into diagnostics and predictive maintenance [48, 49].

3. TLP COI Workshop

The TLP COI Workshop, hosted virtually by NIST on September 24–25, 2024, brought together 48 attendees from industry, academia, and government. Spanning two full days, 22 speakers presented or participated in panels centered on the evolving role of TLP in engineering and industrial applications.

The TLP COI aimed to advance the understanding and application of TLP in engineering and industrial domains through the workshop, with the objective of identifying current challenges, sharing best practices, and fostering collaboration across industry, academia, and government. Expert speakers explored themes such as risk awareness, data accessibility, long-term resource development, and the integration of LLMs into specialized domains. Key sessions highlighted challenges in data sharing, the need for standardized methodologies, and strategies to bridge technical and domain expertise effectively. Interactive sessions encouraged active participation and the exchange of ideas, emphasizing the evolving landscape of TLP tools and opportunities for community-driven innovation.

Each session in the TLP Workshop is summarized, including a summary and takeaway(s) from the perspective of the speaker. The workshop sessions were mostly talks, with a lightning round on day one and a panel session on day two. The talks were either long talks (one hour) or short talks (20-30 minutes). In the lightning round, each speaker had 10 minutes to speak and there was a group Q&A at the end. For the panel session, each speaker presented for five minutes lengthy discussion afterwards.

3.1. Opening: Introduction to TLP

The opening speaker for the event was Michael E. Sharp, of NIST, who provided a brief overview defining TLP while highlighting the need for specialized language interpretation in industrial and scientific settings. In highly contextual or rare event environments, standard NLP will often fail to meet the operator's requirements. In these cases, TLP ensures better cross-domain collaboration between humans and intelligent systems by removing obfuscating language or concept barriers.

3.2. Long Talk: Understanding, Designing & Evaluating Informed NLP Systems

Rachael Sexton, from NIST, provided a more in-depth introduction to TLP methods and tools. She laid out the basic framework that all TLP builds from and highlighted several key methods for development and testing of TLP tools. Her keynote stressed the importance of understanding the assumptions used when developing, implementing, and interpreting a solution using TLP. Establishing clear goals and appropriate performance metrics are necessary prerequisites to successfully implementing a TLP method, she stressed [50].

Key Takeaway: Human language can be more accurately interpreted using approaches from TLP when implementers clearly define their intended outcomes, acknowledge underlying system assumptions, and select appropriate metrics to measure progress.

3.3. Lightning Session: TLP Applications in Government and Industry

The second part of the morning consisted of lighting talks on TLP applications in government and industry.

3.3.1. Improving Product-Service Systems Offerings Through Technical Language Processing

Roberto Sala from the University of Bergamo detailed a case study in which TLP was applied to improve remote troubleshooting in Product-Service Systems (PSS) offerings by leveraging the unstructured data in maintenance reports. The TLP implementation used in the study, which took place in a packaging machine manufacturing company, helped identify design problems and improve maintenance service delivery [51–53]. The study highlights the potential benefits of using TLP in PSS: optimized processes, lower costs, and better customer management.

Key Takeaway: Using TLP in industrial context with focus on improving service delivery and maintenance operations. Knowledge extracted from maintenance reports can lead to lower costs for the company, faster market response times, increased productivity, and more.

3.3.2. TLP for Industrial Innovation: Opportunities for the Factory Floor

Radu Pavel from ARCTOS emphasized the role of TLP in maintenance workflows, particularly for data cleaning and preprocessing as well as hybrid approach which use information across structured and unstructured sources. Hybrid approaches that combine sensor data with operator narratives offer potential for deeper insight during root cause analysis.

Key Takeaway: Combining numeric sensor data with text-based operator input can enhance diagnostics and prognostics, while technical documents—such as proposals, reports, and maintenance logs—offer a pathway for capturing and operationalizing tribal knowledge.

3.3.3. Technical Language Processing to Support the Nuclear Power Industry

Jamie B. Coble from the University of Tennessee provided an overview of TLP in the nuclear sector, where TLP can be used to improve operational efficiency and safety through automating the evaluation and generation of critical documents, improving operational efficiency and safety. Applications include extracting insights from work orders and reports, identifying recurring issues, and facilitating proactive maintenance. Generative Artificial Intelligence (AI) can be used to automate tasks such as creating work orders and drafting licensing documents. By automating routine tasks, these tools can reduce manual workload and enable human experts to focus on more complex aspects of nuclear operations, supporting the goals of safety, reliability, and sustainability.

Key Takeaway: The safety, reliability, and efficiency of complex systems, like nuclear reactors, can be improved by judicious application of TLP techniques.

3.3.4. Utilizing TLP to Get a Holistic Understanding of Freezer Performance

Manjish Naik from John Bean Technologies Corporation presented a use case where TLP was applied to service technicians' comments to generate maintenance troubleshooting recommendations. By combining this unstructured text with time-series sensor data (e.g., vibration monitoring), the system recommended diagnostics for refrigeration equipment. The alert led to identifying root causes—an imbalanced fan wheel and undersized suction line—that impacted energy efficiency and product quality. Corrective actions were implemented, preventing further degradation and production loss.

Key Takeaway: Managing assets through whole lifecycle is a system that involves using different data sources and techniques. TLP is one piece of this puzzle, but it's an important piece of the puzzle.

3.3.5. LLM Quality Assessment & Interests in Design and Manufacturing

Nathan Hertlein, from the Air Force Research Laboratory (AFRL) explored the increasing capabilities of LLMs in the domain of design and manufacturing, where quantitative performance comparisons are challenging due to a lack of domain-specific benchmarks. He and other researchers at the Air Force Research Laboratory are investigating ways to assess the quality of LLM-generated outputs, such as computer-aided design (CAD) files, and evaluating their performance on various use-cases across the design-and-manufacturing workflow. The goal is to measure LLM performance in areas like CAD, manufacturability assessment, and maintenance tasks, as well as robotics action planning, to guide the integration of TLP into control algorithms for real-world hardware.

Key Takeaway: AFRL has been able to use LLMs to generate CAD models by producing scripts for 3D modeling software and synthesize robotic behaviors in state machines [54]. However, prompt design matters: more abstract prompts tend to produce less practical or "crazier" CAD models.

3.4. Long Talk: Generative AI for Engineering Design

Faez Ahmed from Massachusetts Institute of Technology (MIT) spoke on the use of language models in engineering design. His talk was structured in three parts: using vision language models to help designers, designing co-pilots, and complex design spaces.

He began by exploring how vision language models (VLMs) could support early-stage engineering design by helping teams better interpret complex, multi-modal requirements. He introduced *DesignQA*, an open-source benchmark for evaluating multi-modal language models on tasks such as rule extraction, comprehension, and compliance using the Formula SAE design rules [48]). Models like GPT-4o performed best, but overall results highlighted limitations in extracting geometric and dimensional information from engineering drawings.

Additionally, Ahmed proposed the concept of a “Design Co-pilot” using generative models to enhance parametric engineering design. He presented *Design Autocomplete*, where generative models were used to complete or modify parametric CAD models using assembly graphs and parameter embeddings. This approach enabled faster design iterations and produced more diverse alternatives than traditional imputation methods [55]. He also introduced the use of counterfactual reasoning—exploring “what-if” scenarios—to help designers to explore alternative design configurations requiring deep domain-specific expertise. While current generative models are not yet precise enough for every task, he believes rapid advancements is making these tools increasingly relevant to product design workflows.

Finally, Ahmed addressed the difficulty humans face in navigating high-dimensional design problems, such as those with both discrete and continuous components. Using contrastive learning, his team trained joint embedding models on a dataset of 100 million four-bar linkage mechanisms to support data-driven kinematic design [56]. These models support both generative and predictive tasks, including inverse kinematics. He emphasized the importance of structured datasets and domain-specific integration for advancing generative AI capabilities in engineering design.

Key Takeaway: VLMs show early promise in tasks like rule retrieval and compliance checking, but struggle with visual geometry—pointing to a need for better methods to encode and present geometric information in model inputs. Generative models, enhanced by counterfactual reasoning, offer new ways to accelerate design and explore alternatives, but greater precision is needed for broader applicability in real-world design tasks. Solving design problems involving high-dimensional, mixed-variable spaces requires large, structured datasets and novel representations; contrastive learning is one path forward for enabling generative models in such domains. Overall, generative models have incredible potential for transforming the engineering design process, but require careful study to understand their limitations.

3.5. Long Talk: Leveraging AI and NLP for Enhanced Climate and Resilience Research

Juan F. Fung from NIST presented his team’s research for the day’s keynote. His team is focused on developing disaster-resilient infrastructure and communities. By developing and validating AI-assisted human-in-the-loop methods, they aim to make technical resources more accessible and actionable for end-users, such as state and local governments. To support this work, they use TLP to assist—rather than replace—humans in annotating technical documents [57].

A suite of tools, including ALTO, TENOR, and TENOR-GPT, has evolved to support this goal. ALTO applies active learning and LDA for annotation [58]; TENOR improves on this with neural topic models [59]; and TENOR-GPT incorporates large language models to generate label suggestions directly from document content. These tools enable the team to create best practices for AI applications in technical tasks and improve decision-making in areas like community planning.

Key Takeaway: Human-in-the-loop NLP tools enhance climate resilience research by supporting expert-driven annotation of technical documents. Successive tools (ALTO → TENOR → TENOR-GPT) show the progression from topic models to their neural and generative counterparts in annotation workflows. TLP methods help structure technical knowledge essential for building more disaster-resilient systems.

3.6. Short Talk Session: How Can AI Language Models Deliver Business Value?

The first day concluded with a session of short talks from researchers at the University of Western Australia. Melinda Hodkiewicz led the session, tying the talks into a central theme around how AI language models can be developed in ways which provide value to industry. Specifically, the research from Hodkiewicz's lab focuses on practical AI tools for data cleaning, semantic modeling, automated classification, and expert recommendation.

The first short talk was presented by Hodkiewicz, where she discussed integrating AI technologies for enterprise applications with a focus on linked data, LLM challenges, and combining different AI paradigms. She proposed a practical framework for explaining AI to business leaders, in which AI tools are split into three categories: narrow AI generative AI, and logic-based AI—each with distinct roles and benefits.

Key Takeaway: A key message was the importance of shared language, model grounding (e.g., through RAG), and semantic quality assurance as foundations for effective enterprise AI.

3.6.1. Automating Technical Text Classification using AI/NLP.

Michael Stewart explored classifying failure modes from unstructured technical text using LLMs. Early findings emphasized the importance of prompt design—LLMs often invented failure modes unless provided with a predefined list. Fine-tuning offered mixed results, in contrast to the consistency of Narrow AI methods, underlining the challenges in applying generative models to structured classification tasks [14, 20].

Key Takeaway: Prompt engineering is critical when using LLMs for technical classification—unconstrained prompts can lead to hallucinated outputs. While LLMs offer flexibility, traditional Narrow AI approaches may still provide more consistent and reproducible results for specific classification tasks.

3.6.2. Linked Data: The Foundations of Enterprise AI

Caitlin Woods highlighted the role of linked data, or ontologies, as a foundational paradigm for enabling data interoperability and reuse across enterprise systems. Linked data allows organizations to define data concepts in an unambiguous, machine-readable way. This approach supports consistent querying, sharing, and integration of data across diverse software environments [40, 41, 44]. Woods provided examples of companies such as IKEA and the Equinor Sverdrup drilling platform who are already applying linked data strategies. These implementations have shown measurable benefits, including reduced operational costs and improved quality assurance through more consistent and transparent data handling.

Key Takeaway: For long-term AI success, enterprise software and AI systems must operate using a shared, semantically consistent language. In technical domains, this includes grounding AI outputs in physical constraints, which is made possible through linked data systems. These systems support structured reasoning, constraint validation, and interoperable data use. Linked data will be essential to building AI solutions that are robust, audit-able, and aligned with engineering and operational standards.

3.6.3. Garbage In, Garbage Out: Navigating Technical Text Quality – Current Challenges and Future Directions

Tyler Bikaun spoke on navigating technical text quality, highlighting the challenges and opportunities associated with human-generated texts in industry. Poor-quality texts cause information to be misinterpreted by humans, and reduces the performance of NLP applications. This necessarily leads to reduced overall efficiency in language-processing tasks.

Bikaun notes two categories of solution to this problem: narrow AI and generative AI. The narrow AI approach involves “small task-specific language models” improving text quality with the assistance of domain experts. While this method results in consistently high-quality data, it is more costly and resource intensive. On the other hand is generative AI, which offers a comparatively cheap solution to this problem, at the cost of overall quality and reliability. Bikaun suggests that the two approaches could be combined to form a process in which a Generative AI model is fine-tuned by domain experts.

Key Takeaway: The shortenings, typos, and quirks of human-generated text can make for poor results when fed into a language processing algorithm untouched. Results can be improved by translating low-quality text to high-quality text using a narrow AI system trained on expert-annotated text datasets. Generative AI could be used to augment this task, lowering costs.

3.7. Long talk: Standards as Discourse: Technical Language Processing for Standards

Jacob Collard, a computational linguist at NIST, discussed applying NLP methods to convert traditional standards documents into structured, semantic data formats such as knowledge graphs. This shift enables more advanced querying and validation than is possible with document formats like PDF. By modeling standards in this way, the author can extract insights and relationships between elements, such as clauses and terms, and apply this information to tasks like version control, information extraction, and error detection. The goal is to enhance the standards development process and study the language and style used in standards from a scientific perspective, ultimately improving the clarity and usability of these documents.

Working with Eswaran Subrahmanian, Collard showed, with specific examples, how TLP can improve access to complex technical documents and support automated detection of issues like contradictions and circular definitions [60, 61]. The talk sparked questions about how non-experts would navigate linked standards and how this foundational work might help improve the development and use of standards.

Key Takeaway: This approach holds potential for industries such as manufacturing and engineering, where improved access to standards can drive compliance and innovation. By applying TLP, complex documents can be digitized into structured data models representing relationships, making them easier to navigate, validate and apply in real-world contexts.

3.8. Long talk: Building Test Collections for LLMs

Ian Soboroff, of NIST, spoke on the evaluation of LLMs, using experience through the Text REtrieval Conference (TREC), an annual NIST workshop focused on assessing information retrieval systems [62]. The talk looked at the history of evaluation methodologies, including text summarization evaluation and the pyramid method. He noted that many dataset and evaluation practices from retrieval tasks can be adapted to LLMs in both the chat and RAG systems. While many evaluation techniques from information retrieval can be adapted to LLMs, evaluating generative outputs requires new approaches due to their unstructured and variable-length nature and that responses are not ranked. In all cases, context is important for measuring the performance of a system and needs to be considered.

Soboroff also emphasized a major point on LLMs as judge for evaluation. Despite advances in automated evaluation of LLMs (such as using LLMs to judge other LLMs), there is no substitute for human-annotated ground truth to reliability assess LLM performance [63].

Key Takeaway: The current methods for evaluating generated text require significant manual effort, but exploring methods to help automate this process remains a promising active research area.

3.9. Short talk: Text Pre-processing for Predicting Tool Wear

Michael Dawson, a industrial artificial intelligence researcher at NIST, presented their findings on the effect of text pre-processing methods on tool wear predictions, a critical aspect of manufacturing. The study analyzed human observational data from a production environment to predict tool wear using various machine learning models. The results showed that some combinations of text pre-processing steps far outperformed others when used to predict the measured tool wear.

Key Takeaway: Sometimes, the first steps in the language processing pipeline are overlooked. It is important to understand which text pre-processing steps are most apt for your use case, whether by referencing guidance or performing your own evaluations.

3.10. Panel Session: Measuring Impacts and Risk Awareness for TLP

The foundation of improving any tool starts with evaluating the effective impacts of its use. These impacts, both positive and negative, potential and actualized, have associated frequencies, likelihoods, and consequently risks. Understanding these potential risks allows informed and educated decisions about their use. TLP is no different.

In the afternoon, three panelists—AJ Stein (General Services Administration (GSA)), Joshua Gen (LMI), and Udayan Das (Saint Mary's College of California)—discussed the risks and impacts of TLP. The panelists' responses have been paraphrased for brevity.

3.10.1. What are the most common risks you've observed with TLP implementation in industry?

Das emphasized that each document used in training or evaluating a language-processing algorithm is different, so an approach that might work well for a specific document or set of documents may not be applicable to others.

Gen, meanwhile, pointed out that hallucinations by LLMs are a concern, and that they need to be approached differently than other errors. Additionally, data leakage—in which a user's data is shared without their authorization—is a concern that should be considered.

Stein highlighted the problem of "Mechanical Sympathy": when some operators change their behavior to match the machine they're working with, rather than questioning the machine. He added that that the field of cybersecurity doesn't have a consistent language for evaluating data.

3.10.2. What are some common shortcomings you observe regarding communicating TLP risks to stakeholders?

Stein claimed that we need a way for a computer to mediate risks to the user—"cybersecurity for the rest of us."

Das, calling back to Gen's earlier comment, noted that the domains in which language-processing tools are being used aren't necessarily aware of the limitations of those tools. When users in a domain unrelated to language-processing are being offered a new NLP tool, the potential benefit of its use may receive out-sized emphasis compared to its limitations.

Gen agreed, adding that often, it is often the case that many unexpected challenges arise when integrating a new tool into the real-world system, despite interesting and seemingly-streamlined prototypes.

3.10.3. What role do ethical considerations play when evaluating the use of TLP in different industries?

Gen believes that there are many considerations, especially for the federal government. These include: trying to minimize bias, putting guardrails on model outputs, and restricting associations between parts of the output and the source.

Das warned against using off-the-shelf LLMs without understanding their training process and the data sources.

Stein How can one prove to me that the model will do exactly what I need and nothing unexpected? Structured data is incredibly important to even get to the point where you can trace the source and privacy constraints.

Key Takeaway: There is need for mechanisms which ensure output traceability, bias auditing, and control within well-defined guardrails. There is a lack of clear, shared language for assessing cybersecurity risks in data handling. Basic cybersecurity awareness must be ensured across teams, including non-experts, and develop ways for systems to mediate and protect sensitive information.

3.11. Long Talk: Towards Machine-Assisted Reading and Structuring of Scientific and Technology Language

Peter Chung, from the University of Maryland - College Park, presented on the characterization of knowledge graphs through linguistic patterns in technical literature, exploring how machine-assisted reading can support the structuring of domain-specific knowledge. His work addresses the challenge posed by the vast and rapidly growing volume of scientific and engineering publications, and the limitations of current language models that rely on co-occurrence patterns within narrow context windows. Chung introduced two key findings: first, that technical correlations inherent in scientific language—such as name-property relationships for molecules—can be captured and enhanced through data fusion, enriching token representations with domain knowledge. Second, that graph-theoretic methods, combined with ontology rules, can be used to extract structured semantic information directly from unstructured text. This approach holds promise for developing scalable, unsupervised information extraction systems tailored for scientific and technical domains, potentially enabling new capabilities in automated knowledge graph construction, literature review, and technology scouting [64].

Key Takeaway: Integrating domain reasoning, data fusion, and graph-theoretic methods with NLP tools improves unsupervised extraction of structured, domain-specific knowledge from technical text, overcoming the limits of co-occurrence-based models.

3.12. Long Talk: A Modular Platform to Improve AI Usability for Technical Tasks

Philippe Dessauw & Guillaume Sousa Amaral, two NIST researchers, spoke about how the rapid growth of AI, particularly LLMs, has introduced challenges such as generating false information and lacking unified metrics for technical tasks. To address these issues, a modular platform based on RAG has been developed to ensure LLMs generate coherent and contextually relevant responses. The platform's plug-and-play architecture and modular design enable testing, tuning, and replacement of components, providing a robust and efficient solution to enhance AI usability for technical tasks.

Key Takeaway: Design of a modular RAG platform can be used to improve LLM reliability, adaptability, and technical task performance throughout the AI system's lifecycle.

4. TLP and LLM Panel at PHM Society Conference

The TLP and LLM Panel, held during the PHM Society Conference on November 13, 2024, explored the practical applications of TLP and LLMs in the PHM domain. Moderated by Sarah Lukens, the panel brought together a distinguished group of experts who shared insights on how AI-driven tools can enhance PHM workflows, focusing on metrics, best practices, and human-centric implementation.

The panelists were Neil Eklund (Oak Grove Analytics), Michael Sharp, and Hao Huang (GE Vernova). Eklund had presented a tutorial earlier at the conference entitled “LLMs and Multimodal AI for PHM: The Future of Maintenance Intelligence” which introduced approaches for integrating LLMs into PHM workflows, particularly around RAG and multi-modal AI for diagnostics and maintenance optimization. Huang had presented a paper earlier at conference on collaborative root cause analysis using a framework that combines data-driven models with pretrained LLMs to enhance the reliability and accuracy of industrial failure analysis [26].

4.1. Panel Summary

Much of the discussion was centered on the growing synergies between TLP workflows and PHM tasks such as predictive maintenance, failure analysis, and engineering documentation and insights into LLM-assisted solutions for predictive maintenance and engineering documentation. Panelists discussed how LLM-powered solutions can enhance these workflows in supporting decision making and for streamlining technical documentation and unstructured data. While LLMs offer powerful new capabilities, speakers noted that their integration into safety-critical systems requires thoughtful design, robust evaluation, and close collaboration with human experts.

Eklund discussed the computational tradeoffs of using advanced transformer-based models, comparing them to infrastructure systems that must scale with increasing complexity. Sharp urged practitioners to ground AI implementations in well-defined business outcomes and cautioned against over-promising results. Huang illustrated how fine-tuned LLMs can help improve reliability in anomaly detection, but also noted the need for better methods to embed physical principles and expert feedback into the AI loop.

The panel also highlighted the critical role of human-in-the-loop strategies. Human oversight remains essential to ensure the reliability of LLM-based systems, particularly for detecting errors in automated diagnostics, augmenting analyses with domain-specific knowledge, and building trust among engineers and operators.

Panelists addressed cross-cutting challenges such as data standardization, trust, and long-term tool sustainability. Discussions revisited themes from the 2023 PHM panel on generative AI and echoed concerns from the recent NIST TLP workshop, underscoring that engineering rigor—not hype—must guide adoption. A lively audience Q&A session further highlighted the importance of interdisciplinary collaboration and community-driven development to ensure LLM and TLP technologies meet the evolving needs of the industrial sector.

4.2. Key Takeaways from the PHM Panel

- LLMs and TLP systems are increasingly being applied to real-world PHM use cases, including diagnostics, anomaly detection, and engineering documentation.
- Integration of AI into safety-critical workflows must be human-centered, with mechanisms for expert oversight and interpretability.
- Customization of LLMs for industrial domains presents challenges around scalability, cost, and standardization.
- Sustained community engagement and shared metrics are essential for evaluating and maturing these technologies across sectors.

5. Takeaways

A number of key takeaways emerged from the presentations and discussions we witnessed last year, which we summarize here. We have organized these observations into (i) best practices , (ii) methods, and (iii) applications. The major themes in best practices are summarized in Table 1 and described in Section 5.1. Methods are summarized in Table 2 and described in Section 5.2. Section 5.3 covers the different applications presented during the events, summarized in Table 3. Each table summarizes our observations across different speakers, and is cross-referenced with the speaker's last name.

5.1. Best Practices

The sharing of best practices was a significant theme across both events. Broadly, these can be summarized as two best practice principles: (i) follow the engineering design process when creating a TLP system and (ii) center humans in the conception and implementation of that system. Table 1 contains a summary of these practices. The relevance of these practices are described below.

Table 1. Summary of best practices when designing and implementing TLP systems.

Principle	Practice	Speaker(s)
Problem Definition	Thoroughly evaluate the problem that is being solved and the desired outcome that its solution must reach. Be aware of the assumptions that arise during the process.	Sexton, Hodkiewicz, Das
	Understand the characteristics of the available data and quantize it appropriately.	Bikaun, Dawson
Validation	Measure a model's success by comparing its outputs to a non-generated ground truth, rather than another LLM's response. Comparing responses isn't useful if what you want to understand is the model's ability to accurately represent a quantitative measurement.	Soboroff
	Characterize differences between pipelines through parameter variation to verify that the model outputs are representative of the input data.	Dessauw & Amaral
Model Selection & Risk Awareness	Use the class of AI that's appropriate for your use case.	Hodkiewicz
	Understand the risks associated with your use case and the models under consideration.	Stein, Das, Gen
Human-Centered	Design your systems with a human-in-the-loop approach. Experts in their field will, and should, always verify model outputs.	Fung

5.1.1. Use Engineering Design Principles

Engineering design begins with the most important phase of the problem-solving process: defining the problem and the corresponding requirements for a solution [65]. Properly completing this task necessitates that the designer not presuppose any given solution. Failure to do so often shifts the process on its head: the problem is defined according to what the desired solution is able to address. This mistake can result in suboptimal or wholly inaccurate performance and exacerbation of risks not discovered in the problem definition.

Solution-neutrality is the practice of approaching a problem without any preconceived preference towards a particular solution or technology, and is a requirement of a responsible design process. It may be tempting to relax the relax this requirement in the face of the rapid advancements in research and significant financial investments in language processing, where it may be tempting to apply available language models to tasks without fully understanding the problems they are intended to solve.

A key theme of last year's events was proper problem definition. Later stages of the design process, like solution selection and solution validation, were also key themes. Even if a designer accurately models their problem, they must still (i) choose a solution that best meets their stakeholders' requirements and (ii) validate that the chosen solution actually does so in practice [65].

5.1.2. Center Designs around Humans

The second main theme emphasizes the importance of a human-centered approach in designing TLP systems. This approach ensures that the system is designed with the end-users in mind and includes their input and feedback throughout the development process. Key to this approach is incorporating a human-in-the-loop methodology, whereby experts continually verify and validate model outputs. This practice leverages domain expertise to ensure the relevance and accuracy of the TLP system's results.

5.2. Methods

Many methodologies emerged as important or showing potential for future research and development (see Table 2).

5.2.1. Data Quality and Diversity

One common theme was the need to ensure data quality across all stages of language processing (training, input, output). Soboroff cautioned against relying on TLP tools to assess themselves, as this can introduce bias and risk. Integrating diverse and structured data sources, whether linked by design or coincidence, can reduce hidden biases and strengthen model reliability by providing independency, redundant information.

Several speakers discussed the potential to incorporate non-text data type, such as time series data, into LLMs. This is an area of ongoing research with particular promise for applications such as predictive maintenance. Early research suggests that LLMs may be most effective as translators between text and numeric data streams that can be more easily processed by relevant modeling algorithms downstream.

Table 2. Summary of key methods discussed for use in TLP systems.

Method	Description	Speaker(s)
Linked data	Use structured connections and ontologies for interoperability and data reuse.	Collard, Woods
Knowledge graphs	Build and apply knowledge graphs to improve LLM understanding in technical domains.	Chung
RAG	Combine LLMs with external sources to improve accuracy, reduce hallucination.	Ahmed, Dessauw & Amaral, Gen, Hodkiewicz
Fine-tuning & overfitting	Adapt LLMs to specific domains; overfitting can be useful if managed with guardrails.	Eklund
Annotation	Ensure quality training data through careful annotation and active learning.	Fung, Stein
Human-in-the-loop	Involve humans throughout design, training, and deployment for reliability, trust, adaptability.	Fung
Time series data	Integrate time series data with LLMs to enable predictive tasks from heterogeneous data.	PHM Panel

5.2.2. Annotation

Document annotation is critical for developing accurate LLMs. Researchers such as Fung and Stein emphasized the importance of careful annotation practices to ensure high quality and reliable models. Fung's work on human-in-the-loop NLP tools for climate resilience shows how combining active learning with expert annotation can improve model accuracy, while Stein pointed out that careful annotation practices are key for producing models are accurate and reliable.

5.2.3. Linked Data and Knowledge Graphs

Structured approaches like linked data and knowledge graphs help ensure interoperability and reuse across enterprise systems by defining data concepts unambiguously (discussed by Collard and Woods). Chung's work on showed how knowledge graphs can capture domain-specific knowledge and extract semantic information from unstructured text. Using LLMs to build or interact with knowledge graphs can further enhance their performance in technical domains.

5.2.4. Retrieval-Augmented Generation

RAG has gained acceptance for improving LLM performance in specialized tasks by grounding outputs in external knowledge and reducing hallucinations. Many speakers, including Hodkiewicz, Gen, Ahmed, Dessauw, and Amaral described how RAG enhances both accuracy and relevance.

5.2.5. Fine-tuning

Fine-tuning, which is adjusting hyperparameter and reference texts to optimize LLMs for specific tasks is common in academic studies, but has seen less uptake in industry, as noted by Eklund. Barriers include limited expertise, time and resource allocations, as well as uncertainty about its effectiveness and point of diminishing returns. While fine-tuning can improve performance, practical guidance remains limited. Eklund also noted that local overfitting is not always a bad thing if proper guardrails are in place to control outputs. The challenge is balancing between customization and generalization.

5.2.6. Human-in-the-loop Design

A consistent takeaway was the critical importance of human-in-the-loop design and implementation throughout TLP tool development and deployment. Humans serve as auxiliary sensors and safeguards, while also supporting knowledge transfer, training and trust-building between workers and AI systems.

5.3. Applications

A growing body of work is exploring how TLP approaches and LLMs can be applied both in the design of AI systems and in practical industrial applications. The applications, summarized in Table 3, are grouped by domain category and application type. The primary domain categories are maintenance and manufacturing design, but other related technical areas were also covered.

Applications could be categorized as either descriptive or prescriptive and generative. Descriptive applications help users understand or label existing data, while prescriptive and generative enable decision-making, automation, and the creation of new designs or content. Descriptive applications often include from improving data quality in maintenance logs to assisting engineers in generative design tasks or diagnostics.

As shown in Table 3, the applications discussed at last year's events spanned multiple industrial domains. Maintenance applications discussed during the workshop and panel included using TLP to support root cause analysis, automate document generation, and extract insights from unstructured logs. In engineering design, applications include using language models to assist with tasks such as CAD scripting, parametric completion, and creative design exploration. Annotation and labeling tools

Table 3. Summary of industrial applications of LLMs.

Domain	Application Type	Task	Speaker(s)
Maintenance	Descriptive	Cleaning, annotation and classification of unstructured maintenance data	Pavel, Bikaun, Stewart, Coble
	Descriptive	Failure mode classification	Stewart, Bikaun
	Descriptive	Root cause analysis combining numeric and textual data	Pavel, Naik
	Descriptive	Work order trend analysis in nuclear maintenance logs	Coble
	Descriptive	Operator log processing for tool wear prediction	Dawson
	Descriptive	Linked data representations of maintenance concepts	Woods
	Generative	Remote troubleshooting using unstructured maintenance reports	Sala
	Generative	Task automation (e.g., work order and license drafting)	Coble
	Generative	Diagnostic recommendations combining technician comments and sensor data	Naik
Design	Generative	LLM-driven CAD scripting	Hertlein, Ahmed
	Generative	Manufacturability and robotics planning support	Hertlein
	Generative	CAD completion	Ahmed
	Generative	Exploring Design modifications or alternatives	Ahmed
Other	Descriptive	Topic label suggestion for climate resilience documents	Fung
	Descriptive	Active learning for technical report labeling	Fung
	Descriptive	Knowledge graph representation of technical chemistry documentation	Chung
	Descriptive	Knowledge graph representation of standards	Collard

also play an important role, enabling domain experts to more efficiently work with large collections of technical documents. The applications summarized here represent only those covered at the events and are by no means an exhaustive list.

6. Conclusion

The 2024 TLP COI Workshop at NIST and the TLP and LLM panel at the 2024 PHM Society Conference highlighted progress made as well as identifying and articulating challenges in advancing TLP across industry, academia and government. Speakers explored critical topics spanning methods, best practices and applications. Looking ahead, the TLP community has the opportunity to continue cross-collaboration, share lessons learned, develop practical guidelines, and help educate the workforce on how to use these technologies effectively in ways which provide value. Future work should focus on addressing complex challenges in ways that align with the specific data, applications, and precision requirements of industry.

References

[1] Brundage MP, Sexton R, Hodkiewicz M, Dima A, Lukens S (2021) Technical Language Processing: Unlocking Maintenance Knowledge. *Manufacturing Letters* 27:42–46. DOI:10/grd9dk. Available at https://www.sciencedirect.com/science/article/pii/S2213846320301668

[2] Dima A, Lukens S, Hodkiewicz M, Sexton R, Brundage MP (2021) Adapting Natural Language Processing for Technical Text. *Applied AI Letters* DOI:10/gkchch

[3] Lukens S, Ali A (2023) Evaluating the Performance of ChatGPT in the Automation of Maintenance Recommendations for Prognostics and Health Management. *Annual Conference of the PHM Society* (PHM Society), Vol. 15. DOI:10/g9rtzs

[4] Brundage M, Weiss B, Pellegrino J (2020) Summary Report: Standards Requirements Gathering Workshop for Natural Language Analysis (National Institute of Standards and Technology, Gaithersburg, MD), NIST AMS 100-30. DOI:10/g9rtx7

[5] TLP COI (2024) Technical Language Processing Community of Interest 2024 Meeting. Available at https://www.nist.gov/news-events/events/2024/09/technical-language-processing-community-interest-2024-meeting.

[6] PHM Society (2024) Panel 08: TLP and LLM: Enhancing PHM Capabilities. Available at https://phm2024.phmsociety.org/wp-content/uploads/sites/14/2024/12/Panel-08-TLP-LLM-PHM-Panel-2024.pdf.

[7] Navinchandran M, Sharp ME, Brundage MP, Sexton R (2022) Discovering critical KPI factors from natural language in maintenance work orders. *Journal of Intelligent Manufacturing* 33:1859–1877. DOI:10/gtjzb8

[8] Seale M, Hines A, Nabholz G, Ruvinsky A, Eslinger O, Rigoni N, Vega-Maisonet L (2019) Approaches for Using Machine Learning Algorithms with Large Label Sets for Rotorcraft Maintenance. *2019 IEEE Aerospace Conference* (IEEE), pp 1–8.

[9] Trilla A, Mijatovic N, Vilasis-Cardona X (2022) Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting. *International Journal of Prognostics and Health Management* 13(2). DOI:10/hbc7p8

[10] Bikaun T, Hodkiewicz M (2021) Semi-automated Estimation of Reliability Measures from MaintenanceWork Order Records. *PHM Society European Conference* (PHM Society), Vol. 6, pp 9–9. DOI:10/hbc7rt

[11] Hansen B, Coleman C, Zhang Y, Seale M (2020) Text Classification and Tagging of United States Army Ground Vehicle Fault Descriptions in Support of Data-Driven Prognostics. *Annual Conference of the PHM Society* (PHM Society), Vol. 12, pp 8–8. DOI:10/hbc7rf

[12] Dixit S, Mulwad V, Saxena A (2021) Extracting Semantics from Maintenance Records. *CoRR* abs/2108.05454. DOI:10/hbc7pf

[13] Sexton R, Hodkiewicz M, Brundage MP (2019) Categorization Errors for Data Entry in Maintenance Work-Orders. *Annual Conference of the PHM Society* (PHM Society), Vol. 11. DOI:10/gtj4hz

[14] Stewart M, Hodkiewicz M, Li S (2023) Large Language Models for Failure Mode Classification: An Investigation. DOI:10/qgvj.

[15] Iyer N, Virani N, Yang Z, Saxena A (2022) Mixed Initiative Approach for Reliable Tagging of Maintenance Records with Machine Learning. *Annual Conference of the PHM Society* (PHM Society), Vol. 14. DOI:10/hbc7q8

[16] Sexton R, Brundage MP, Hoffman M, Morris KC (2017) Hybrid datafication of maintenance logs from AI-assisted human tags. *International Conference on Big Data (Big Data)* (IEEE), pp 1769–1777. DOI:10/gtj4hf

[17] Nandyala AV, Lukens S, Rathod S, Agarwal P (2021) Evaluating Word Representations in a Technical Language Processing Pipeline. *PHM Society European Conference* (PHM Society), Vol. 6, pp 17–17. DOI:10/hbc7qw

[18] Hodkiewicz M, Ho MTW (2016) Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering* 22(2):146–163. DOI:10/hbc7rc

[19] Lukens S, Naik M, Saetia K, Hu X (2019) Best Practices Framework for Improving Maintenance Data Quality to Enable Asset Performance Analytics. *Annual Conference of the PHM Society* (PHM Society), Vol. 11. DOI:10/g9rtx6

[20] Stewart M, Hodkiewicz M, Liu W, French T (2022) MWO2KG and Echidna: Constructing and exploring knowledge graphs from maintenance data. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* DOI:10/grd3kp

[21] Hodkiewicz MR, Batsioudis Z, Radomiljac T, Ho MT (2017) Why autonomous assets are good for reliability–the impact of 'operator-related component'failures on heavy mobile equipment reliability. *Annual Conference of the PHM Society* (PHM Society), Vol. 9. DOI:10/hbc7s8

[22] Hodkiewicz M, Klüwer JW, Woods C, Smoker T, Low E (2021) An ontology for reasoning over engineering textual data stored in FMEA spreadsheet tables. *Computers in Industry* 131:103496. DOI:10/hbc7pj

[23] Rajpathak D, Cafeo J (2021) A Semantic Similarity Model to Compare Heterogeneous Data Sources to Augment Engineering Data with New Failure modes in Automotive Industry. *PHM Society European Conference* (PHM Society), Vol. 6, pp 10–10. DOI:10/hbc7qn

[24] Yang C, Shen W, Chen Q, Gunay B (2018) A practical solution for HVAC prognostics: Failure mode and effects analysis in building maintenance. *Journal of Building Engineering* 15:26–32. DOI:10/hbc7p5

[25] Löwenmark K, Taal C, Schnabel S, Liwicki M, Sandin F (2022) Technical Language Supervision for Intelligent Fault Diagnosis in Process Industry. *International Journal of Prognostics and Health Management* 13(2). DOI:10/gq7gnc

[26] Huang H, Shah T, Karigiannis J, Evans S (2024) Physics and Data Collaborative Root Cause Analysis: Integrating Pretrained Large Language Models and Data-Driven AI for Trustworthy Asset Health Management. *Annual Conference of the PHM Society* (PHM Society), Vol. 16. DOI:10/qgt5

[27] Lukens S, McCabe LH, Gen J, Ali A (2024) Large Language Model Agents as Prognostics and Health Management Copilots. *Annual Conference of the PHM Society* (PHM Society), Vol. 16. DOI:10/hbc7q2

[28] Pau D, Tarquini I, Iannitelli M, Allegorico C (2021) Algorithmically Exploiting the Knowledge Accumulated in Textual Domains for Technical Support. *PHM Society European Conference* (PHM Society), Vol. 6, pp 12–12. DOI:10/hbc7qs

[29] Ferdousi R, Hossain MA, Yang C, Saddik AE (2024) DefectTwin: When LLM Meets Digital Twin for Railway Defect Inspection. DOI:10/qgvm.

[30] Bastos P, Lopes I, Pires L (2012) A Maintenance Prediction System using Data Mining Techniques. *World Congress on Engineering 2012* (International Association of Engineers), Vol. 3, pp 1448–1453. Available at https://www.iaeng.org/publication/WCE2012/WCE2012_pp1448-1453.pdf.

[31] Bokinsky H, McKenzie A, Bayoumi A, McCaslin R, Patterson A, Matthews M, Schmidley J, Eisner L (2013) Application of Natural Language Processing Techniques to Marine V-22 Maintenance Data for Populating a CBM-Oriented Database. Available at https://sc.edu/study/colleges_schools/engineering_and_computing/docs/research/predictive_maintenance_pdfs/ahs2013bokinsky.pdf.

[32] Pires F, Leitão P, Moreira AP, Ahmad B (2023) Reinforcement learning based trustworthy recommendation model for digital twin-driven decision-support in manufacturing systems. *Computers in Industry* 148:103884. DOI:10/hbc7qp

[33] Ansari F (2020) Cost-based text understanding to improve maintenance knowledge intelligence in manufacturing enterprises. *Computers & Industrial Engineering* 141:106319. DOI:10/grfcgq

[34] Ansari F, Glawar R, Nemeth T (2019) PriMa: a prescriptive maintenance model for cyber-physical production systems. *International Journal of Computer Integrated Manufacturing* 32(4-5):482–503. DOI:10/gm52tp

[35] Siddharth L, Blessing L, Luo J (2022) Natural language processing in-and-for design research. *Design Science* 8:e21. DOI:10/hbc7qb

[36] Ito A, Hagström M, Bokrantz J, Skoogh A, Nawcki M, Gandhi K, Bergsjö D, Bärring M (2022) Improved root cause analysis supporting resilient production systems. *Journal of Manufacturing Systems* 64:468–478. DOI:10/gqpjtx

[37] Brundage MP, Kulvatunyou B, Ademujimi T, Rakshith B (2017) Smart Manufacturing Through a Framework for a Knowledge-Based Diagnosis System. *Proceedings of the ASME 2017 12th International Manufacturing Science and Engineering Conference* (ASME), Vol. 3. DOI:10/hbc7rs

[38] Papageorgiou K, Theodosiou T, Rapti A, Papageorgiou EI, Dimitriou N, Tzovaras D, Margetis G (2022) A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing. *Frontiers in Manufacturing Technology* 2:972712. DOI:10/hbc7qt

[39] Karray MH, Ameri F, Hodkiewicz M, Louge T (2019) ROMAIN: Towards a BFO compliant reference ontology for industrial maintenance. *Applied Ontology* 14(2):155–177. DOI:10/ggvhj3

[40] Woods C, Selway M, Bikaun T, Stumptner M, Hodkiewicz M (2023) An ontology for maintenance activities and its application to data quality. *Semantic Web* 15(2):319–352. DOI:10/hbc7pk

[41] Woods C, Hodkiewicz M, French T (2024) Semantic Quality Assurance of Industrial Maintenance Procedures. *IEEE Access* 12:122029–122046. DOI:10/qgt4

[42] Vidyaratne L, Lee XY, Kumar A, Watanabe T, Farahat A, Gupta C (2024) Generating Troubleshooting Trees for Industrial Equipment using Large Language Models (LLM). *International Conference on Prognostics and Health Management (ICPHM)* (IEEE), pp 116–125. DOI:10/hbc7p7

[43] Rajpathak DG (2013) An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain. *Computers in Industry* 64(5):565–580. DOI:10/f4x6gv

[44] Woods C, French T, Hodkiewicz M, Bikaun T (2023) An ontology for maintenance procedure documentation. *Applied Ontology* 18(2):169–206. DOI:10/qgtx

[45] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H (2023) Retrieval-Augmented Generation for Large Language Models: A Survey. DOI:10/gtxbg9.

[46] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih Wt, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, eds Larochelle H, Ranzato M,

Hadsell R, Balcan M, Lin H, Vol. 33, pp 9459–9474. Available at https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[47] Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Larson J (2024) From Local to Global: A Graph RAG Approach to Query-Focused Summarization. DOI:10/qgvr.

[48] Doris AC, Grandi D, Tomich R, Alam MF, Ataei M, Cheong H, Ahmed F (2024) DesignQA: A Multimodal Benchmark for Evaluating Large Language Models' Understanding of Engineering Documentation. *Journal of Computing and Information Science in Engineering* 25(2). DOI:10/g89tc3

[49] Löwenmark K (2025) *Technical Language Supervision and Agentic AI for Condition Monitoring*. Ph.D. thesis. Luleå University of Technology, . Available at https://www.diva-portal.org/smash/get/diva2:1950819/FULLTEXT02.pdf.

[50] Sexton R (2022) Text as Data: the Road to Technical Language Processing!. Available at https://tlp-coi.github.io/text-data-course/home.html.

[51] Sala R, Pirola F, Pezzotta G, Cavalieri S (2023) Improvement of maintenance-based Product-Service System offering through field data: a case study. *Production & Manufacturing Research* 11(1). DOI:10/g9rtzw

[52] Sala R, Pirola F, Pezzotta G, Cavalieri S (2022) NLP-based insights discovery for industrial asset and service improvement: an analysis of maintenance reports. *IFAC-PapersOnLine* 55(2):522–527. DOI:10/hbc7qg

[53] Sala R, Pirola F, Dovere E, Cavalieri S (2019) A Dual Perspective Workflow to Improve Data Collection for Maintenance Delivery: An Industrial Case Study. *Advances in Production Management Systems. Production Management for the Factory of the Future*, eds Ameri F, Stecke KE, von Cieminski G, Kiritsis D, pp 485–492. DOI:10/hbc7qh

[54] Swick B, Donegan S, Gillman A, Groeber M (2024) Human Planning of Robot Actions through LLM-guided State Machine Synthesis. *International Conference on Robot and Human Interactive Communication (ROMAN)* (IEEE), pp 430–437. DOI:10/g9rt2j

[55] Regenwetter L, Abu Obaideh Y, Ahmed F (2023) Counterfactuals for Design: A Model-Agnostic Method for Design Recommendations. *Proceedings of the ASME 2023 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (ASME), Vol. 3A. DOI:10/hbc7qk

[56] Heyrani Nobari A, Srivastava A, Gutfreund D, Ahmed F (2022) LINKS: A Dataset of a Hundred Million Planar Linkage Mechanisms for Data-Driven Kinematic Design. *Proceedings of the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (ASME), Vol. 3A. DOI:10/hbc7rd

[57] Fung JF, Li Z, Stephens DK, Mao A, Goel P, Walpole E, Dima A, Boyd-Graber JL (2024) Human-In-The-Loop Technical Document Annotation: Developing and Validating a System to Provide Machine-Assistance for Domain-Specific Text Analysis (National Institute of Standards and Technology, Gaithersburg, MD), NIST TN 2287. DOI:10/g9rtz9

[58] Poursabzi-Sangdeh F, Boyd-Graber J, Findlater L, Seppi K (2016) ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, eds Erk K, Smith NA (ACL), Vol. 1, pp 1158–1169. DOI:10/g9rtxw

[59] Li Z, Mao A, Stephens D, Goel P, Walpole E, Dima A, Fung J, Boyd-Graber J (2024) Improving the TENOR of Labeling: Re-evaluating Topic Models for Content Analysis. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, eds Graham Y, Purver M (ACL), Vol. 1, pp 840–859.

[60] Collard J, de Paiva V, Subrahmanian E (2024) Mathematical entities: Corpora and benchmarks. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, eds Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N (ELRA and ICCL, Torino, Italia), pp 11080–11089. Available at https://aclanthology.org/2024.lrec-main.966.pdf.

[61] Monarch I, Collard J, Shin S, Subrahmanian E, Bhat TN, Sriram RD (2022) Making Semantic Structures Explicit: Developing and Evaluating Tools and Techniques to Support Understanding of Large Cybersecurity Corpora (National Institute of Standards and Technology, Gaithersburg, MD), NIST IR 8414. Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933620.

[62] Voorhees EM, Harman DK (eds) (2005) *TREC* (The MIT Press).

[63] Soboroff I (2025) Don't Use LLMs to Make Relevance Judgments. *Information Retrieval Research Journal* 1(1):29–46. DOI:10/hbc7p9

[64] O'Ryan C, Hayes KD, VanGessel FG, Doherty RM, Wilson W, Fischer J, Boukouvalas Z, Chung PW (2025) An Automated Approach for Domain-Specific Knowledge Graph Generation-Graph Measures and Characterization. *Journal of Chemical Information and Modeling* DOI:10/qgtw

[65] Dieter GE, Schmidt LC (2013) *Engineering Design* (McGraw-Hill), 5th Ed.

Appendix A 2024 TLP COI Workshop Organizers

Table 4. Organizers of the 2024 TLP COI Workshop.

Organizer	**Affiliation**
Michael E. Sharp	*NIST*
Sarah Lukens	*LMI Consulting LLC*
Rachael Sexton	*NIST*
Alden Dima	*NIST*
Michael Dawson	*NIST*
Michael Brundage	*Applied Research Laboratory for Intelligence and Security (ARLIS)*

Appendix B Acronyms

AFRL: Air Force Research Laboratory
AI: Artificial Intelligence
ARLIS: University of Maryland Applied Research Laboratory for Intelligence and Security
CAD: Computer-Aided Design
COI: Community of Interest
DOI: Document Object Identifier
FMEA: Failure Mode and Effects Analysis
GSA: General Services Administration
LDA: Latent Dirichlet Allocation
LLM: Large Language Model
MD: Maryland
MIT: Massachusetts Institute of Technology
NIST: National Institute of Standards and Technology
NLP: Natural Language Processing
ORCID: Open Researcher and Contributor Identifier
PDF: Portable Document Format
PHM: Prognostics & Health Management
PSS: Product-Service Systems
RAG: Retrieval-Augmented Generation
SAE: Society of Automotive Engineers
SP: Special Publication
TLP: Technical Language Processing
TREC: Text REtrieval Conference
U.S.: United States
VLM: Vision Language Model