LLM-Enabled Real Time Training Content Curation to Enhance Performance

John Carney, Nancy Belmont, James King

MARi LLC

Alexandria, Virginia

john.carney@mari.com, nancy.belmont@mari.com, james.king@mari.com

Dr. John Stamper, Christine Kwon, Shruti Srinivasan

Carnegie Mellon University

Pittsburgh, Pennsylvania

jstamper@andrew.cmu.edu, ckwon2@andrew.cmu.edu, shrutisr@andrew.cmu.edu

Jeff Withington

Google LLC

Mountain View, California

jwith@google.com

ABSTRACT

Service members' skill and knowledge gaps can result in significant safety and readiness gaps. Recent advancements in Large Language Models and Multimodal Models can be leveraged to enhance the effectiveness of ongoing readiness initiatives such as Ready Relevant Learning (RRL). Detailed multimodal content (e.g. text, video, mixed reality) can be rapidly tagged and organized to ensure continuous alignment with ever-changing technical publications to close skill and knowledge gaps.

For example, while inspecting an aircraft for discrepancies, AI-enabled content curation could allow an aircraft maintainer to click on a technical publication step and see a detailed video tutorial that showcases common discrepancies on the relevant part of the aircraft. In this case, the crucial information gets to the right person in the right modality to improve performance.

Text-only technical publications that support repair and maintenance can limit understanding and do not necessarily connect with younger generations. Multimedia tutorials are generally more effective materials than text for instructing hands-on procedures; however, keeping multimedia content up-to-date is currently time and labor intensive as frequent updates to official technical publications occur.

We outline an automated, AI-enabled schematic linkage between a given task, e.g. "inspect the nose of an F/A-18", and its related content objects in various media formats. Users can toggle between video and text instructions while executing a complex procedure to ensure the necessary detail is delivered to the user, avoiding costly errors. Automated linkage of an official procedure step and its aligned content objects also enables flagging a section of content for update or human review if the official technical publication is changed, ensuring that content remains up-to-date. In tests that leverage car and bicycle maintenance procedures and tutorial videos, we demonstrate the ability to automatically tag and maintain multimedia tutorial content using AI methods with 79 - 98% effectiveness, measured by various metrics detailed in the paper.

ABOUT THE AUTHORS

John Carney, Principal Investigator, is the Founder & CEO of MARi, an artificial intelligence (AI) web-based Software as a Service (SaaS) for academic and workplace skills. John has 30 years of experience in the field of learning technologies, and is also co-founder and CEO of Carney, Inc., a training and performance acceleration company.

John Stamper is an Associate Professor at the Human-Computer Interaction Institute at Carnegie Mellon University. He is also the Technical Director of the Pittsburgh Science of Learning Center DataShop. His primary areas of research include Educational Data Mining and Intelligent Tutoring Systems.

Christine Kwon is a PhD student at the Human-Computer Interaction Institute at Carnegie Mellon University. Her current areas of research include educational technologies and learning sciences.

Nancy Belmont is Vice President at MARi. She drives growth and commercialization effortswith 30+ years of organizational development and business experience.

James King is the Product Manager at MARi and a graduate of the Carnegie Mellon University School of Computer Science with a Masters in Education Technology and Applied Learning Sciences.

Shruti Srinivasan is an undergraduate student at Carnegie Mellon University studying Statistics and Machine Learning and Artificial Intelligence. She is a teaching assistant in Computer Science and Artificial Intelligence and a research assistant in Human-Computer Interaction.

Jeff Withington is a Customer Engineer at Google focused on delivering solutions to the DoD. He is a United States Navy Veteran.

INTRODUCTION

Currently, training in the armed forces is dominated by instructor-led schoolhouses. Official technical publications, which are a standing order, are used as a reference by various service members (e.g. aircraft maintainers) as they do their jobs. Problematically, it is not feasible for schoolhouse training to build service member fluency in every procedure that they need to perform; even if there were time to do this, official technical publications and their associated maintenance and repair procedures change over time. While service members can reference the official technical publication for a refresher on procedures that are new, or for which they've experienced skill decay, tech pubs lack the detail needed to show readers the nuance of performing a complex maintenance or repair procedure [Mayer, 2013]. Multimedia content (e.g. videos or mixed reality content), on the other hand, can depict in detail the nuanced tasks that a maintainer must perform on an aircraft. Multimedia instruction has also yielded significantly higher learner performance in manual tasks when compared against text instruction [Donkor, 2020].

The benefits of affording warfighters access to tech pub-aligned multimedia content can be seen in the high costs of maintenance errors. F/A-18 corrosion, which is greatly impacted by the effectiveness of maintenance procedures on the aircraft, cost the Navy over $2 billion between 2017 - 2020, per The Navy Times. By enhancing F/A-18 maintainer performance, billions of dollars could be saved while increasing aircraft uptime and safety. F/A-18 maintainers' effectiveness at handling new and complex procedures could be greatly improved with on-demand access to detailed multimedia content aligned with official manuals. While official tech pubs are required to be used by maintainers, they are not detailed enough to instruct; tech pubs are necessary but insufficient.

While the performance benefits brought by tech pub-aligned detailed multimedia tutorials are evident, keeping multimedia content up-to-date with ever-changing technical publications has traditionally been laborious and costly. Every time a technical publication is updated, all of its aligned pieces of training content need to be manually inspected to ensure that they reflect the updated tech pub. We outline below a novel method of semi-automating alignment between official technical publications and multimedia content. When a new piece of content (e.g. a video) is ingested, the Step-by-Step Tutorial Engine for Performance Readiness (STEPR) system automatically uses AI to detect whether each required tech pub step is present and captures the timecodes of each present step. A video

missing a required tech pub step is automatically flagged by the STEPR system to ensure accuracy of performance support content made available to warfighters. Further, if a technical publication is updated, all aligned pieces of multimedia content can be automatically processed to check which sections require updates to be aligned with the tech pub. In this way, time is saved by directing human reviewer attention only to those sections of content impacted by the particular update to the tech pub. Schematic linkage of technical publications and their associated multimedia content enables leading-edge video intelligence and natural language processing AI methods to continuously ensure that multimedia content stays aligned with the current version of the tech pub, removing tedious, error-prone maintenance labor and providing warfighters with reliable content to support on-the-job performance.

BACKGROUND & RELATED WORK

Comparing Video-Based and Text-Based Learning for High-Risk Tasks

YouTube-like videos have become a primary source of learning for many, instigating many to compare video-based learning to text-based learning [List 2018, Tarchi et al 2021], particularly for complex procedural tasks. Surgical procedures serve as a prime example of tasks that significantly benefit from video-based learning [Buch et al 2014, Routh et al 2023, Srinivasa et al 2020]. In particular, prior research investigated differences between video and text-based learning to determine which modality of learning promotes higher student learning in sensitive tasks, such as basic laparoscopic suturing and other clinical procedures, in which video-based learning was shown to be more effective than text-based learning in these contexts [Routh et al 2023]. Additionally, Sonnenfeld et al. examined training procedures in electronic and distance learning approaches for flight crew training. In fact, they found that video-based learning was generally effective for procedure-based training in providing flight crew trainees with interactive opportunities and formal training content [Sonnenfeld et al 2021]. As video-based learning can effectively instruct and convey information that is difficult to teach through standard text, it is crucial to understand how to implement and provide effective video-based learning for high-risk tasks that are dependent on a correct sequence of key steps.

Procedural videos vary widely in content, especially in the hierarchy of steps and key moments for learning particular tasks. Hence, there is a noticeable growth in interest in enhancing the searchability and indexing of instructional videos by extracting key informational features [Zala et al 2023]. This form of information retrieval can be executed through methods such as video-moment retrieval, hierarchical step extraction, and step summarizations [Zala et al 2023]. Our work employs a similar approach by extracting systematized schemas that organize the hierarchy and context of task steps verified through an empirical measure of comparison. As current research is extending these content-based feature extraction methods to instructional videos for ubiquitous task-based learning, we aim to enhance task-based learning support for more complex and sensitive tasks.

LLM-Based & Video Intelligence-Based Extraction Methods

Currently, Large Language Models (LLMs) are showing their immense potential in efficiently carrying out diverse NLP tasks [Ampel et al 2023, Bang et al 2023, Brown et al 2020, Kojima et al 2022, Liu et al 2023, Purwar and Sundar 2023, Wei et al 2022, Zhu et al 2023]. However, we have yet to completely rely on LLMs as little is known about the ability of LLMs to verify their content output [Zhang & Gao 2023], especially in dissecting and validating complex media content. Popular LLM-generated content validation methods, such as “chain of thought” prompting and Hierarchical Step-by-Step prompting, include the use of prompting to verify content subcomponents [Wei et al 2022, Zhang & Gao 2023]. However, verification methods using prompt chaining may not be entirely efficient for extracting and verifying multiple ordered steps of a task, particularly when the task data is unstructured. In fact, more recent studies tested LLM-based methods for text annotations, text summarization, and organized information extractions, which also include representations of extracted information in knowledge graphs [Carta et al 2023, Goel et al 2023]. In light of high-risk cases that may benefit from these LLM-based solutions, prior studies have employed LLM-based approaches to extract organized annotations and informational features from, for instance, unstructured clinical data [Agrawal et al 2022, Goel et al 2023]. More recent studies have shown success in extracting task steps from video narrations using LLM-based support [Shang et al 2023]. However, we have yet to determine how these LLM-based solutions can extract and verify hierarchical task steps from such high-risk procedures, especially in the form of media content.

The detection of key steps within instructional videos is difficult, as they rely on both the chronological and temporal ordering of steps in accomplishing a task. Additionally, the accurate localization of these steps is dependent on their alignment with a validated sequence of instructions when completing a task. Text transcripts of instructional tasks are conventionally compared to a standardized list of instructions [Logeswaren et al 2023, Malmaud et al 2015]. However, recent studies on the alignment of video demonstrated steps to a "gold standard" manual are not completely reliant on narration-based feature extractions. Rather, they employ a joint method that evaluates text-based extractions in conjunction with visual feature detectors [Alayrac et al 2016, Malmaud et al 2015]. These particular studies aim to bridge the gap between visual and linguistic variability of steps from instructional videos [Alayrac et al 2016, Mavroudi et al 2023].

According to Tang et al (2023), LLMs can analyze video content in two primary ways, which are text extraction from videos for response generation and a combined approach in which LLMs are combined with visual models to finetune and create a cohesive model for video content analysis. In fact, with the exponential growth and potential of LLMs, recent studies have shown the potential to finetune LLM-based textual extraction with visual detection models, including video and image captioning models [Shang et al 2023, Tang et al 2023]. However, these combined approaches of the LLM-and-visual model have not yet proven effective for high-risk and sensitive procedures that require important attention to step details. Hence, our work aims to use a similar combined approach to align complex multimedia content with technical publications.

REPOSITORY ARCHITECTURE

The Knowledge Object (KO) data structures depicted below enable alignment between multiple media objects such as source-of-truth technical publications and aligned multimedia content. It is critical that the official technical publication is maintained as the single source of truth and that other attached media are aligned with the tech pub. To accomplish this, we define a KO as a procedure and all its related tasks, descriptions, and aligned pieces of content. The procedure's tasks, descriptions, and other details are pulled from the technical publication into a nested data structure that contains information describing the KO and the required steps or actions that are involved in the given process (e.g. inspect aircraft left-forward fuselage for discrepancies). Figure 1 shows the JSON structure that stores official steps using the organizational hierarchy of the written tech pub. Any change made by human officials to an official tech pub will be automatically updated in the tech pub's data structure.


```
{
    "id": "7.1.3",
    "title": "Volvo XC60 Drive Preparation",
    "manual": "MDART Version 2023 1.2.4",
    "steps": {
        "Tire Replacement":
        {"sub-steps":
            {
            "1": "Apply parking brake.",
            "2": "Remove the spare tire from the trunk.",
            "3": "Use wheel chocks to block the wheels opposite the one being changed.",
            "4": "Loosen the lug nuts on the wheel to be changed.",
            "5": "Raise the vehicle with a jack.",
            "6": "Remove the lug nuts and the wheel.",
            "7": "Install the new tire.",
            "8": "Screw the lug nuts back on.",
            "9": "Lower the vehicle.",
            "10": "Tighten the lug nuts."
            },
            "id": "III-7-2"
        },
        "Oil Change":
        {"sub-steps":
            {
            "1": "Park the vehicle on a level surface.",
            "2": "Turn off the engine.",
            "3": "Open the hood.",
            "4": "Locate the oil dipstick.",
            "5": "Pull the dipstick out and wipe it clean.",
            "6": "Reinsert the dipstick and pull it out again.",
            "7": "Check the oil level."
        },
        "id": "III-7-1"
        }
```


Figure. 1 . The above showcases a sample official procedure taken from the auto maintenance field. Related multimedia segments can reference each procedure step in a complementary schema.

Each piece of multimedia content, including the example video with a data structure depicted in Figure 2, has its data nested in a schema that specifies the official tech pub and step(s) aligned with the video. Each tech pub step demonstrated in the video is aligned with a distinct video moment that contains start and stop timecodes. The

linkage of each moment in a video to the tech pub step demonstrated allows specific moments of videos to be flagged when a tech pub is updated, streamlining the video update process.


```
{
    "video-id": "1",
    "knowledge-object-id": "7.1.3",
    "knowledge-object-components": {
        "steps": {
            "Oil Change": {
                "sub-steps": {
                    "1": "Park the vehicle on a level surface.",
                    "2": "Turn off the engine.",
                    "3": "Open the hood.",
                    "4": "Locate the oil dipstick.",
                    "5": "Pull the dipstick out and wipe it clean.",
                    "6": "Reinsert the dipstick and pull it out again.",
                    "7": "Check the oil level."

                }

            }
    }
    },
    "video-title": "Volvo XC60 Oil Change",
    "video-description": "This video demonstrates how to change the oil in a Volvo XC60.",
    "video-url": "https://www.MARi.com/LearningActivities/watch?v=VolvoXC60OilChange",
    "video-length": "03:30",
    "video-moments": {
        "0": {
            "start-time": "00:00",
            "end-time": "00:30",
            "sub-step": "1"

    },
        "1": {
            "start-time": "00:30",
            "end-time": "01:00",
            "sub-step": "2"
```


Figure. 2 . The above showcases a sample tutorial video linked with several official procedure steps.

If a tech pub step or procedure is changed, aligned media objects are automatically flagged for human review. Further, the specific moments in a video impacted by a manual change are flagged to increase the focus and efficiency of human review of flagged videos.

The below diagram outlines the lifecycle of a piece of multimedia content in the STEPR system.

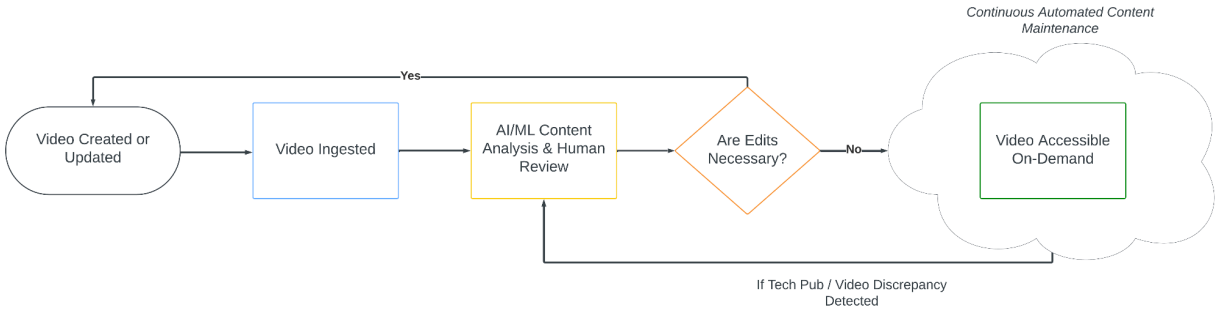

Figure. 3 . A flowchart that outlines the maintenance lifecycle of content in the STEPR system.

USE CASES

On-Demand Performance Support for Warfighters

Automated content curation and storage based in the Knowledge Object schema facilitates streamlined on-demand access to vetted performance support content. The ability to link a technical publication step and related moment in a tutorial video, for example, can enable an aircraft maintainer to click on a publication step and see a detailed tutorial video segment demonstrating how to execute the step. For example, an aircraft maintainer inspecting a form-in-place (FIP) seal could click on the "inspect FIP seal" step in the tech pub and see a video that contrasts common discrepancies present on a FIP seal with a FIP seal that should pass an inspection. As evidenced by past studies, multimedia instruction has yielded significantly higher learner performance in manual tasks when compared against text instruction [Donkor, 2020]. Detailed performance support multimedia materials linked with technical publications stand to reduce maintenance error rates, ultimately improving aircraft safety and uptime.

Process Automation to Save Warfighter Time and Improve Readiness

Keeping multimedia content up-to-date with changing technical publications typically requires significant human maintenance labor; each time a technical publication is updated, all aligned content needs to be reviewed and/or updated. Reliance on laborious processes to update content can cause knowledge repositories to quickly become out of date, rendering the knowledge repository unusable. The outlined AI process of schematically linking each step of a technical publication with its aligned content segment(s) enables automated flagging of content segments, e.g. video moments, impacted by a technical publication change. For example, a person tasked with keeping training content up-to-date can review the two minutes of a video impacted by a technical publication change instead of the hour-long video. In the below UI, the human-in-the-loop's attention is drawn to the key differences between the old version of the tech publication and the updated version. Content updates can be requested or content can be left in the video with a warning, depending on the severity of an update to a tech pub.

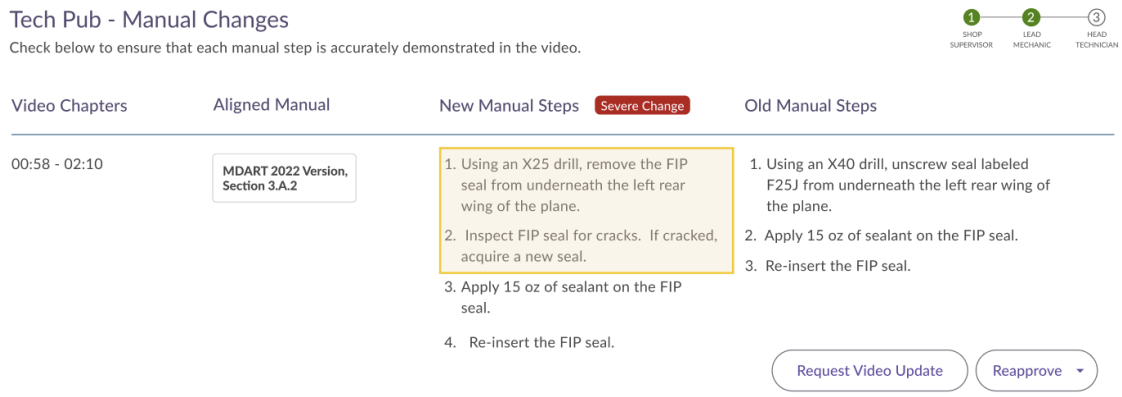

Figure. 4 A sample user interface that depicts what the human-in-the-loop sees when an official procedure step linked with a tutorial video has changed.

Research & Development With Non-Classified Content

We collected two video datasets on two distinct auto-maintenance tasks: replacing a flat tire on a car and repairing a chain on a bicycle. The first video dataset consisted of 23 videos on flat tire replacement of a car, and the second video dataset consisted of 20 videos on the task of bike chain repair. For each dataset, YouTube-generated transcripts were used as input data for the LLM-based pipeline. We selected car and bike maintenance procedures because of the large quantity of publicly available tutorial videos of these procedures and the well-defined, step-by-step structure of car and bike maintenance procedures.

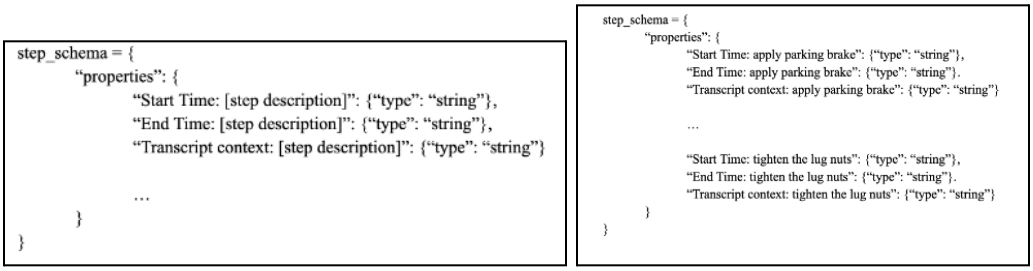

Fig. 5. The left figure is the outline of the schema input, which consists of three features: the start time and end times of the demonstrated step and the transcript context used to determine the time interval of each step occurrence. The right figure is the schema input used to extract key task steps for each video on replacing a flat tire on a car.

Table 1. This table outlines the unique features listed in the properties for each task step of the input schema. Each feature is presented with its type and definition.

Feature Definition in Input Schema		
Feature Name	**Feature Type**	**Feature Definition**
Start Time: [step description]	string	Starting timestamp on first narrated mention or on-screen action of step
End Time: [step description]	string	Ending timestamp on last narrated mention or on-screen action of step
Transcript context: [step description]	string	Portion of text from the video narration on the full occurrence of a step

Table 2. This table lists 12 official manual steps for flat tire replacement on a car, used to identify and verify the retrieved steps within the schema extraction.

Manual Steps on Flat Tire Replacement on a Car	
Step Number	**Manual Step**
1	Apply parking brake
2	Remove the spare tire from the car
3	Use wheel chocks to block the wheels opposite of the wheel you're changing
4	Loosen the lug nuts from the tire
5	Loosen the jack
6	Use the jack to lift up the car
7	Remove lug nuts
8	Remove flat tire
9	Place the spare tire
10	Screw on the lug nuts

11	Use the jack to lower the car
12	Tighten the lug nuts

Methodology
We employed LangChain, a language model integration framework powered by Large Language Models (LLMs) to develop and extricate organized schemas from any text source [Chase 2023]. We specifically used LangChain in conjunction with GPT 3.5 and GPT 4.0 models to create and extract schemas that align the retrieved steps with affirmed key steps from any official manual. Using the LangChain framework, we created an input schema, labeled "step_schema," shown in Figure 5, that lists fundamental key steps on a procedural task, in this case, car tire replacement, from an official manual within an organized knowledge metastructure. Within the schema, we listed the desired extracted features for each step, which include the starting timestamp of when a step is first mentioned, the ending timestamp when a step is no longer mentioned, and the transcript context used to extract each step occurrence from a video narration. The extracted features for each step of a task are also explained in Table 1. Figure 5 also presents the specific schema input used in our approach to extract the key steps listed in Table 2 from videos on car flat tire replacement.

In providing an input schema with key task steps, this LLM-based pipeline can structure an output schema isomorphic to the input schema. By systematizing task steps and their features within an organized schema, crucial step information is easily accessible and directly aligned with authenticated task steps from a verified source.

Analyses
To evaluate the performance of the LLM-based pipeline, three human evaluators conducted a cross-check, comparing the time intervals of each step extracted by the LLM-based pipeline with the actual time intervals of the corresponding task steps in each video. This detailed process involved verifying that the extracted steps accurately aligned with real-time occurrences of task steps in each video. This analysis was necessary in identifying and segmenting task steps within the video content.

Below, we detail the definitions of our method's Identification Rate, Precision, and Recall used to evaluate the effectiveness of the content-tech pub alignment algorithm.

- **Identification Rate:** The proportion of procedure steps actually demonstrated in each video identified by the LLM-based pipeline out of all steps actually demonstrated in each video.
- **Precision:** The proportion of procedure steps identified timestamps out of all identified task steps by the LLM-based pipeline.
    - An identified step from the LLM-based pipeline is considered to be correct if its extracted start and end timestamps overlap with the time interval of actual step demonstrated in the corresponding video.
- **Recall:** The proportion of true positive cases identified as positive by the LLM-based pipeline.
    - We defined a true positive to be an extracted step by the LLM-based pipeline that is also a real step occurrence in a video.

Results
The LLM-based pipeline demonstrated strong performance across each dataset. Specifically, for the video dataset on flat tire replacement on a car, the summary metrics include an average identification rate of 98% for the steps, with a precision of 86% and a recall of approximately 91%. These summary metrics are presented in Figure 8. For the video dataset on bike chain repair, the metrics include an average identification of 95% for the steps, with a precision of 80% and a recall of about 91%. Figure 6 illustrates the detailed extraction results for each video in both datasets. Individual metrics for each video in each dataset are represented in Figures 6 and 7.

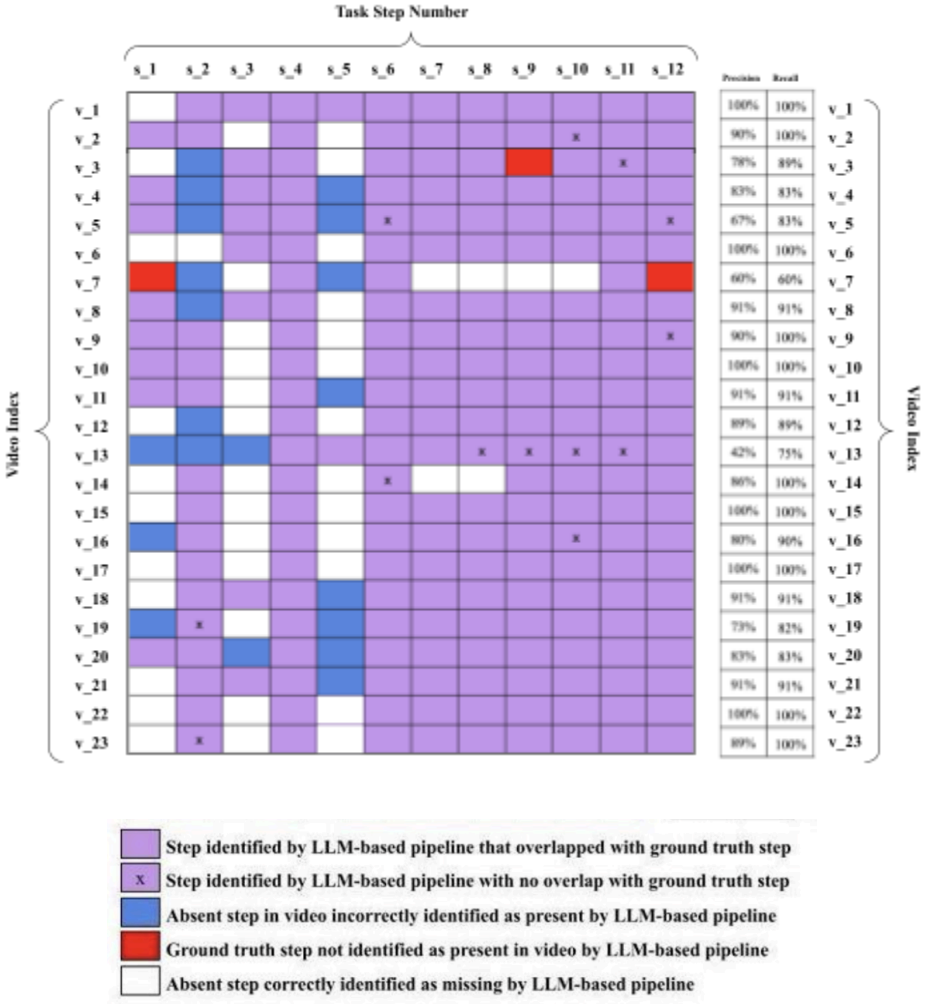

Figure 6. This grid displays the extracted and identified steps by the LLM-based pipeline for each video in the dataset on replacing a flat tire on a car. The legend explains the different results that can occur when comparing the outputs of the LLM-based pipeline to ground truth steps in each video.

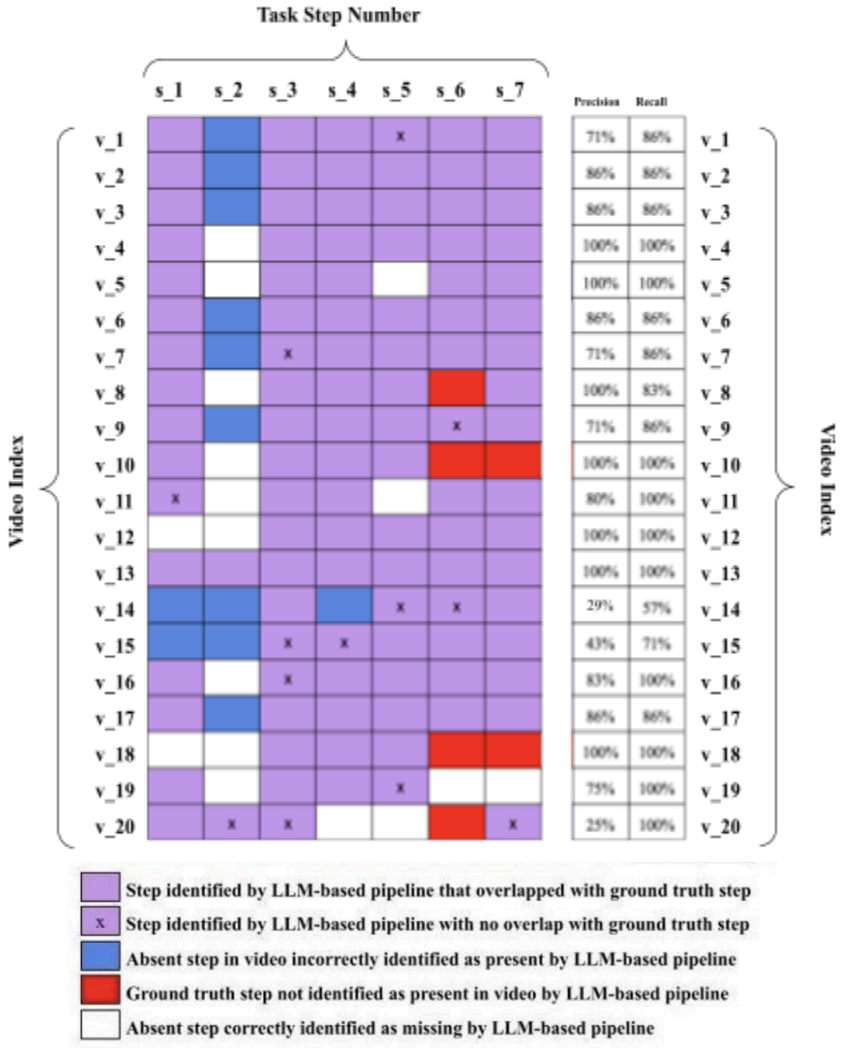

Figure 7. This grid displays the extracted and identified steps by the LLM-based pipeline for each video in the dataset on repairing a bike chain.

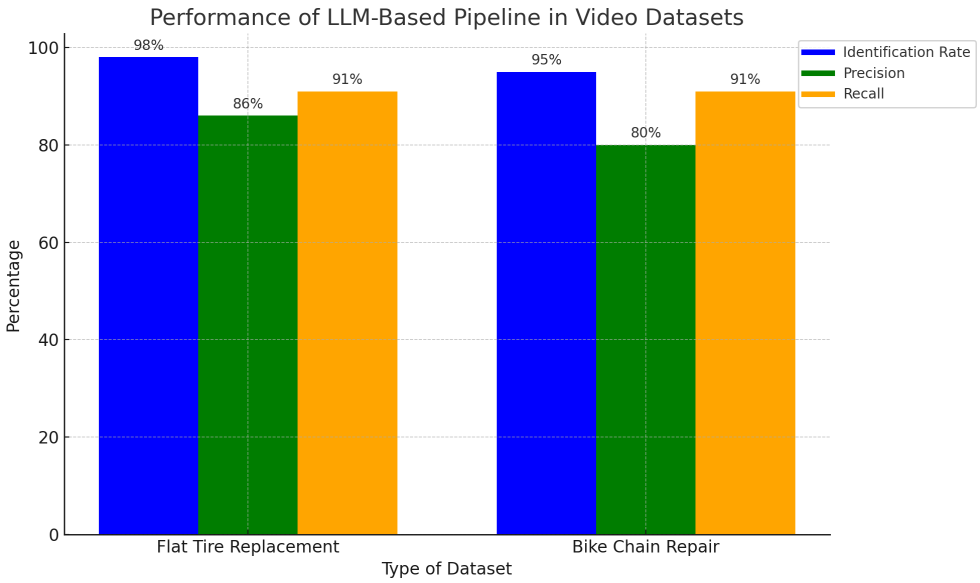

Figure 8. The bar chart illustrates the performance of the LLM-based pipeline across two video datasets. For flat tire replacement, the pipeline achieved an average identification rate of approximately 98%, a precision of 86%, and a recall of 91%. For bike chain repair, the identification rate was about 95%, with a precision of 80% and a recall of 91%.

Preliminary Video Intelligence Results

Our team is currently exploring our ability to leverage emerging video intelligence methods to detect required procedure steps in videos that do not contain transcripts. In this case, only actions and objects detected on screen are compared against the list of procedure steps when evaluating whether a required step is present. We removed audio from 8 of the car tire replacement videos detailed above and leveraged Google's multimodal Gemini 1.5 Pro foundation model to detect whether each video contains each required step; we similarly prompted the model to output the start and stop times of any present step. This approach attained an average Recall of 84% and stands to be enhanced from future fine-tuning.

FUTURE APPLICATIONS

Overcoming Present Limitations

The STEPR methods outlined in this paper should be tested with aircraft maintenance video content and official technical publications to more effectively gauge the accuracy of the methods in the aircraft maintenance context. Further, an operational pilot will be necessary to evaluate the impact of the STEPR system on metrics such as aircraft discrepancy identification, uptime, and safety.

While the use of AI to verify content accuracy stands to both save time and catch content inaccuracies, AI systems will also need scrutiny to ensure that their own accuracy is maintained. The human-in-the-loop methodology employed by the STEPR system is one approach to allow a human subject matter expert to fact-check the output of an AI system before a given piece of content is approved or flagged. In the high-stakes context of the DoD, maintaining human oversight and control over AI systems is essential to the safety and performance of servicemembers who leverage AI systems to do their jobs. The STEPR system could have explainability features layered on in the future to enable the human-in-the-loop to understand why the software views a given required step as present or missing, enabling more informed decision making.

Real Time AI Assistance & Tutoring

As AI assistants and tutors are rapidly adopted throughout the DoD and in other mission critical environments, AI and human-in-the-loop vetting of content referenced during Retrieval-Augmented Generation (RAG) is vital to build

AI system accuracy. The above-described technology lays the foundation for development of an On-Demand AI Maintenance Tutor and Assistant that can process a video feed and warfighter questions in near-real time to provide performance support. For example, a camera on a maintainer's helmet could detect whether corrosion is present on a given part of the aircraft and guide the maintainer with best practices for responding to the particular type and location of the corrosion, improving aircraft safety and uptime. The relevant technical publication and its aligned pieces of content could be referenced as trusted sources via a Retrieval-Augmented Generation (RAG) pipeline to ensure that detailed, current support is provided to warfighters on demand.

Auto-Generation of Training Content, e.g. Courses and Simulations

Automatic tagging of multimedia content and technical publications also lays a technological foundation for automated generation of training content. For example, videos of an aircraft coupled with their associated technical publications could be leveraged to automatically generate an immersive training on aircraft repair and maintenance. Further, coupling these videos and tech pubs with maintainer proficiency data could enable personalized content generation to target and close the skill gaps of an individual warfighter.

DISCUSSION & CONCLUSION

In the DoD context, delivery of accurate, up-to-date content in the right modality to the right service member leads to improved safety, up-time, and readiness. Service members can only rely on the accuracy of multimedia content used to support performance if the content is continuously aligned with technical publications. The technical publication is the source of truth vetted by human experts; associated multimedia content is continuously kept aligned with the technical publication by the STEPR content maintenance system we outline in this paper. In the Marine Corps, for example, the T&R Manual is already the source-of-truth document used to train Marines. The STEPR system would enable each section of the T&R Manual to be schematically linked with multimedia content (e.g. tutorial videos, AR simulations, powerpoint presentations) so that crucial details not available in text-only publications get to trainees to enhance warfighter performance.

As previously outlined, the STEPR system is built to handle ever-changing technical publications. Each time a given procedure step in a technical publication is updated, the multimedia content associated with the procedure step is automatically flagged for review by STEPR. The human-in-the-loop can determine whether and how to update content aligned with the updated technical publication step. As referenced in the future applications section, generative AI could be used to automatically update a given piece of content after its aligned tech pub step(s) is updated as foundation model and fine-tuning capabilities improve.

We believe that generative AI may have implications for the creation, consumption, and evaluation of a servicemember's training. Given the speed and accuracy at which generative AI can compare actions against instructions and generate media, future studies should explore what an end-to-end generative AI training stack looks like.

REFERENCES

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Alayrac, J. B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., & Lacoste-Julien, S. (2016). Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4575-4583).

Ampel, B. M., Yang, C. H., Hu, J., & Chen, H. (2023). Large Language Models for Conducting Advanced Text Analytics Information Systems Research. *arXiv preprint arXiv:2312.17278*.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Buch, S. V., Treschow, F. P., Svendsen, J. B., & Worm, B. S. (2014). Video-or text-based e-learning when teaching clinical procedures? A randomized controlled trial. *Advances in Medical Education and Practice*, 257-262.

Carta, S., Giuliani, A., Piano, L., Podda, A. S., Pompianu, L., & Tiddia, S. G. (2023). Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

Chase, H. LangChain LLM App Development Framework. https://langchain.com/ (accessed August 1, 2023)

Correll, Diana Stancy. (2021). F/A-18 corrosion maintenance doesn't consistently meet Navy and Marine Corps standards. Navy Times.

Donkor, F. (2010). The comparative instructional effectiveness of print-based and video-based instructional materials for teaching practical skills at a distance. International review of research in open and distributed learning, 11(1), 96-116.

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., ... & Feder, A. (2023, December). Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)* (pp. 82-100). PMLR.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.

List, A. (2018). Strategies for comprehending and integrating texts and videos. *Learning and Instruction*, *57*, 34-46.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35.

Logeswaran, L., Sohn, S., Jang, Y., Lee, M., & Lee, H. (2023). Unsupervised task graph generation from instructional video transcripts. *arXiv preprint arXiv:2302.09173*.

Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A., & Murphy, K. (2015). What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*.

Mavroudi, E., Afouras, T., & Torresani, L. (2023). Learning to ground instructional articles in videos through narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15201-15213).

Purwar, A., & Sundar, R. (2023, October). Keyword Augmented Retrieval: Novel framework for Information Retrieval integrated with speech interface. In *Proceedings of the Third International Conference on AI-ML Systems* (pp. 1-5).

Routh, D., Rao, P. P., Sharma, A., & Arunjeet, K. K. (2023). To Compare the Effectiveness of Traditional Textbook-Based Learning with Video-Based Teaching for Basic Laparoscopic Suturing Skills Training-A Randomized Controlled Trial. *Medical Journal of Dr. DY Patil University*.

Shang, C., Tran, E., Narasimhan, M., Subramanian, S., Klein, D., & Darrell, T. (2023). LUSE: Using LLMs for Unsupervised Step Extraction in Instructional Videos.

Sonnenfeld, N., Nguyen, B., Boesser, C. T., & Jentsch, F. (2021). Modern Practices for Flightcrew Training of Procedural Knowledge. In *21st International Symposium on Aviation Psychology* (p. 303).

Srinivasa, K., Chen, Y., & Henning, M. A. (2020). The role of online videos in teaching procedural skills to post-graduate medical learners: A systematic narrative review. *Medical Teacher*, *42*(6), 689-697.

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., ... & Xu, C. (2023). Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Tarchi, C., Zaccoletti, S., & Mason, L. (2021). Learning from text, video, or subtitles: A comparative analysis. *Computers & Education*, *160*, 104034.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Zhang, X., & Gao, W. (2023). Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., ... & Wen, J. R. (2023). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.