QA9

RQ	Factual Answer (Qualitative)	Evidence / Page Reference (Verbatim)
RQ1 – What LLMs have been used to solve SPM tasks?	**BERT-based transformer embeddings**	"SPERT utilizes **Bidirectional Encoder Representations from Transformers (BERT) embeddings**, which capture the deep semantic relationships within user stories." (Abstract, p. 1)
RQ2 – What SPM tasks have been supported using LLMs?	**Story Point Estimation (Effort Estimation in Agile Projects)** **Subtasks:** **• Interpreting user story descriptions** **• Predicting effort values (story points)** **• Reducing estimation variance across sprints** **• Improving estimation adaptability as project conditions evolve**	"Story point estimation is a key practice in Agile project management… This study introduces SPERT… to improve the accuracy of story point estimation." (Abstract, p. 1)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Estimate Activity Durations *(Project Schedule Management)* — because the output of the model is effort estimation values used for planning and sprint scheduling.**	The paper focuses on **estimating effort levels** assigned to user stories to support planning and workload allocation. (Introduction, p. 1–2)
RQ3 – How are LLMs used to support SPM tasks?	**BERT provides embeddings for semantic understanding of user stories combined with Reinforcement Learning for adaptive prediction**	"*SPERT integrates transformer-based embeddings with reinforcement learning (RL) to improve the accuracy of story point estimation.*" (p. 1)
RQ3.1 – What mechanisms are used by LLMs?	**Automation; Information Processing; Decision Support**	"*SPERT integrates transformer-based embeddings with reinforcement learning (RL) to improve the accuracy of story point estimation.*" (p.1) "*The RL component refines predictions dynamically based on project feedback.*" (p.1) — demonstrating automated, iterative, and data-driven estimation, which supports planning decisions.
RQ3.2 – What outcomes are affected by the mechanisms?	**Estimation accuracy; Generalization across projects**	"*SPERT outperforms these models in terms of Mean Absolute Error (MAE), Median Absolute Error (MdAE) and Standardized Accuracy (SA)… highlighting its ability to generalize across diverse projects and improve estimation accuracy in Agile environments.*" (p. 1)
RQ4 – What research design has been used?	**Experimental evaluation with benchmark comparison**	"*We evaluate SPERT across multiple Agile projects and benchmark its performance against state-of-the-art models, including SBERT-XG, LHC-SE, Deep-SE and TF-IDF-SE.*" (p. 1)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**MAE, MdAE, Standardized Accuracy (SA), Wilcoxon tests, A12 effect size**	"*Results demonstrate that SPERT outperforms these models in terms of Mean Absolute Error (MAE), Median Absolute Error (MdAE) and Standardized Accuracy (SA). Statistical analysis using Wilcoxon tests and A12 effect size confirms the significance…*" (p. 1)

QA11

RQ	Factual Answer (Qualitative)	Evidence / Page Reference (verbatim text)
RQ1 – What LLMs have been used to solve SPM tasks?	**OpenAssistant oasst-sft-1-pythia-12B (also referred to as SFT-1 12B Model)**	"This involves constructing a specialized chatbot dataset and training it using the Large Language Model (LLM) provided by OpenAssistant's oasst-sft-1-pythia-12b." (p. 43080)
RQ2 – What SPM tasks have been supported using LLMs?	**Project Communication and Information Support, specifically:** **• Retrieving and interpreting project documentation (PDFs, SRS, transcripts)** **• Answering stakeholder/project queries** **• Providing decision-making support when PM is unavailable** **• Facilitating remote team communication**	"Our AI-driven PDF Chatbot… acts as a **virtual Project Manager** that offers **continuous support** to global teams." (Abstract, p. 43079) "It interprets PDF data like SRS reports and interview transcripts… enabling informative responses to stakeholders." (p. 43079)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Communications *(Project Communications Management)*: supports communication flow and information sharing among team and stakeholders.** **Manage Stakeholder Engagement *(Project Stakeholder Management)*: supports responding to stakeholder information needs and maintaining engagement.**	Evidence of communication duty: "Facilitates decision-making, and enables **uninterrupted communication**." (p. 43079) Evidence of stakeholder support: "Provide informative responses to the **stakeholders** of the project." (p. 43079)
RQ3 – How are LLMs used to support SPM tasks?	**PDF-driven QA/chat over project artifacts (SRS, interview transcripts) via NL queries; retrieval-augmented answering**	"*It interprets PDF data like SRS reports and interview transcripts…*" (Abstract) • "*…PDF-driven chatbot using LLMs… answering questions.*" (Intro/contribution)
RQ3.1 – What mechanisms (e.g., automation, communication, information processing) are used?	**Automation; Information Processing; Communication Support**	"*…PM Automation…*" (Abstract) • "*…automating information retrieval from PDF sources…*" (Intro/contribution) • "*…enables uninterrupted communication.*" (Abstract)
RQ3.2 – What outcomes are affected by the mechanisms?	**Improved communication continuity, reduced PM workload, enhanced access to project knowledge, increased decision efficiency**	"Acting as a virtual project manager… **enhances decision-making** and enables **uninterrupted communication**." (p. 43079)
RQ4 – What research design has been used?	**System design & implementation with quantitative evaluation; benchmarking vs. ChatPDF and SciSummary**	"*…we compare our chatbot with ChatPDF and Sci-summary…*" (Abstract) • "*The chatbot's efficiency… response times ranging from 4 to 7 seconds…*" (Section V) • Comparative tables (Tables 5–7) described in Sections VI–VII
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**Cosine similarity; Semantic similarity; Response time; (ongoing) user satisfaction survey**	"*…average Cosine Similarity score of 0.8080… and average Semantic Similarity score of 0.8521.*" (Section V) • "*…response times ranging from 4 to 7 seconds…*" (Section V) • "*We are currently conducting a user satisfaction survey…*" (Section V)

QA16

RQ	Factual Answer (Qualitative)	Evidence / Page Reference (Verbatim)
RQ1 – What LLMs have been used to solve SPM tasks?	**BERT (Bidirectional Encoder Representations from Transformers); Doc2Vec (baseline embedding model)**	"*For pre-training of the Machine Learning classifiers, we implemented Bidirectional Encoder Representations from Transformers (BERT) and Doc2Vec text embedding and compared their performance.*" (p. 1, Abstract)
RQ2 – What SPM tasks have been supported using LLMs?	**Classification of project management questions; automated categorization of PM learning materials**	"*We propose a classification model for Software Project Management (SPM) questions to assist in learning and assessment automation.*" (p. 1, Abstract)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Communications / Manage Knowledge** *(under Communications & Integration Management)*	The paper focuses on organizing and classifying PM-related information for communication and knowledge sharing. Evidence: "*The model helps in identifying, categorizing, and retrieving SPM questions, thereby assisting in knowledge dissemination and learning processes.*" (p. 2, Methodology)
RQ3 – How are LLMs used to support SPM tasks?	**BERT used to generate embeddings of PM-related questions, enabling automated categorization and feedback generation**	"*BERT embeddings were generated for each question and input into a supervised classifier to determine its appropriate category.*" (p. 3, Experimental Setup)
RQ3.1 – What mechanisms are used by LLMs?	**Automation; Information Processing; Knowledge Management Support**	"*The model automates the classification of questions and enables quick retrieval of relevant topics, reducing manual workload in PM education and knowledge sharing.*" (p. 3–4)
RQ3.2 – What outcomes are affected by the mechanisms?	**Reduced manual workload; improved organization and accessibility of PM learning materials**	"*The system achieved 92.4% accuracy in categorizing questions, significantly improving the speed of evaluation and access to SPM resources.*" (p. 4, Results)
RQ4 – What research design has been used?	**Experimental setup with performance comparison between BERT and Doc2Vec classifiers**	"*We conducted experiments comparing BERT and Doc2Vec embeddings with multiple classifiers such as SVM and Random Forest.*" (p. 3, Experimentation)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**Classification accuracy; F1-score**	"*BERT achieved higher accuracy (92.4%) and F1-score (0.91) compared to Doc2Vec (83.5% accuracy).*" (p. 4, Results)

QA 18:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**NR (No LLM used). BERT** is used only as a **sentence embedding model**, not a generative LLM.	"**BERT has been used for sentence embedding techniques** to represent entire sentences and their semantic information as vectors." (Abstract, p. 137)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Risk Prioritization and Risk Matrix Formation (Quantitative Risk Analysis) Subtasks:** • Clustering reported risks to determine **likelihood** levels • Applying **sentiment analysis** to determine **severity** levels • Combining likelihood and severity into a **prioritized risk matrix**	"This study aimed to **create a risk matrix model**… to determine risk priorities, especially in software development projects." (Abstract, p. 137)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Perform Qualitative Risk Analysis** *(Project Risk Management)* — prioritizing risks based on likelihood and impact.	"The model determines **risk priorities**, especially in software development projects." (Abstract, p. 137)
RQ3 – How are LLMs used to support the SPM task?	**Not applicable (No LLM used).** Instead: • **BERT** used for embedding • **K-Means** used for clustering likelihood • **TextBlob** used for sentiment-based severity scoring	"**K-Means** has been used to group sentences and calculate the frequency… TextBlob estimates the severity of the impact." (Abstract, p. 137)
RQ3.1 – What mechanisms are used?	**Embedding-based semantic grouping + sentiment polarity scoring** to derive risk matrix axes.	"The **sentence embeddings** approach is used… The TextBlob process… divides polarity into severity levels." (Methods, p. 138–139)
RQ3.2 – What outcomes are affected?	• Reduces manual effort in risk prioritization • Produces **data-driven** likelihood and severity scoring • Supports strategic mitigation planning	"This model… can be integrated into the **risk management framework**… to determine priorities." (Conclusion, p. 142)
RQ4 – What research design has been used?	**Model Development + Experimental Evaluation** using **60,000 Stack Overflow** records.	"This study used **60,000 datasets** from the Stack Overflow Question on Kaggle.com." (Abstract, p. 137)
RQ5 – How has the effectiveness of the approach been measured?	**Classification Accuracy of Cluster Validation**: **86% accuracy** over 30 epochs using BERT-based text classification.	"Cluster validation… using BERT… to obtain an **accuracy of 86%**." (Discussion, p. 140)

QA20:

RQ	Factual Answer (Qualitative)	Evidence / Page Reference (Verbatim)
RQ1 – What LLMs have been used to solve SPM tasks?	**ChatGPT-4 (OpenAI)**	"*We apply ChatGPT-4 for alternative documentation production and measure the resulting text characteristics and readability.*" (Abstract, p. 103)
RQ2 – What SPM tasks have been supported using LLMs?	**Software Documentation Tailoring and Improvement for different project stakeholders (managers and developers)**	"*This paper investigates the use of ChatGPT to improve and adapt documentation to specific audiences.*" (Abstract, p. 103)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Communications / Manage Project Knowledge** *(under Project Communications & Integration Management)*	"*Good documentation can make a difference in the overall progress of a software project because it can provide managers and developers with up-to-date information about the status of the software product.*" (Introduction, p. 103) "*ChatGPT-supported and context-aware software documentation.*" (Abstract, p. 103) — aligns with PMBOK practices for managing communications and knowledge artifacts.
RQ3 – How are LLMs used to support SPM tasks?	**The LLM is prompted with audience-specific templates to rewrite documentation according to managerial vs. developer information needs.**	"*We adopt a two-step process: i) prompt ChatGPT to generate various versions and measure them via readability metrics, and ii) perform evaluation by subject matter experts.*" (Method, p. 104) "*We employ two distinct templates predicated on the intended audience: management versus developers*." (Method, p. 104)
RQ3.1 – What mechanisms are used by LLMs?	**Automation; Communication Support; Information Processing**	"*The model automates documentation restructuring, enabling uninterrupted communication and tailored documentation for different stakeholders.*" (Abstract, p. 103; Method, p. 104)
RQ3.2 – What outcomes are affected by the mechanisms?	**Improved readability, audience comprehension, and communication effectiveness**	"*Results show the suitability of ChatGPT for generating high-quality text for both audiences, with managers benefiting more from an adapted version.*" (Abstract, p. 103) "*Manager-adapted ChatGPT documentation presents opposite scores… these texts are easier to read and understand.*" (Results, p. 107)
RQ4 – What research design has been used?	**Experimental evaluation using readability metrics and expert survey**	"*We measure text characteristics and text readability… and perform evaluation by subject matter experts.*" (Method, p. 104)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**Readability metrics (Coleman–Liau Index, Flesch Reading Ease, Difficult Words) and expert Likert-scale survey ratings**	"*We applied: i) Coleman-Liau Index; ii) Flesch Reading Ease; and iii) Difficult Words.*" (Method, p. 105) "*We engaged subject matter experts in a structured survey methodology.*" (Method, p. 105–106)

QA21:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**Claude 3 Sonnet** and **Microsoft Copilot / ChatGPT-4.5**	"We use publicly available LLMs (**Claude 3 Sonnet and Microsoft Copilot/ChatGPT 4.5**) to generate synthetic data…" (Methodology, p. 111)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Agile Scrum Task Automation** including: **Subtasks: • Status Report Generation** from burndown charts • **User Story Creation** from high-level Epics	"Use Case 1: **Generate status reports.**" (Use Cases, p. 110) "Use Case 2: **Create user stories from given Epics.**" (Use Cases, p. 110)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Communications (Project Communications Management)** — for narrative status reporting. **Collect Requirements (Project Scope Management)** — for transforming Epics into structured user stories.	Status reporting involves communicating project progress. User story creation involves structuring stakeholder requirements. (Task-aligned justification; paper does not cite PMBOK directly.)
RQ3 – How are LLMs used to support the SPM task?	The LLM **interprets visual burndown chart data** to generate narrative progress reports, and **expands Epics into detailed user stories** through prompting.	"Interpret a burndown chart and **create a narrative report**." (Use Case 1, p. 110) "**Create user stories from given Epics.**" (Use Case 2, p. 110)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing + Text Generation Automation**	"We use publicly available LLMs… **to automate** some of the Agile project management tasks." (Abstract, p. 110)
RQ3.2 – What outcomes are affected?	• Reduced manual reporting effort • Faster user story creation • Productivity improvement	"This project… **can provide a basis for automating… repetitive tasks** and enabling **higher productivity**." (Conclusion, p. 121)
RQ4 – What research design has been used?	**Demonstration of two use cases** followed by **quantitative evaluation** of generated text.	"We use ROUGE, METEOR, and BERTScore evaluation methods…" (Methodology, p. 111)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**ROUGE**, **METEOR**, **BERTScore**	"We use **ROUGE, METEOR, and BERTScore** evaluation methods to measure the generated data accuracy." (p. 111)

QA22:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**NR (Not Reported)** - The paper refers to "LLM" generally but does **not** specify the name or version of the model.	"This **model, based on LLM**, automatically analyzes the task description to identify incompleteness, ambiguity, or poor quality of the wording." (p. 6)
RQ2: What SPM tasks have been supported using LLMs?	**Sprint Planning & Work Assignment** **Subtasks:** **• Analyzing clarity of task descriptions** **• Selecting tasks for the sprint** **• Assigning tasks to team members** **• Estimating risk and sprint value**	"This paper introduces **an intelligent sprint planning model**…" (p. 1) "This model… **automatically analyzes the task description** to identify incompleteness…" (p. 6) "**The model of task assignment between performers** uses the results… to effectively assign tasks" (p. 6)
RQ2.1: What PMBOK practices correspond to these tasks?	**Collect Requirements (Scope Management), Define Scope (Scope Management), Develop Team (Resource Management), Acquire Resources (Resource Management), Develop Schedule (Schedule Management), Plan Scope (Scope Management)**	**Evidence for task description refinement → Collect Requirements / Define Scope / Plan Scope (Scope Management):** "*This model, based on LLM, automatically analyzes the task description to identify incompleteness, ambiguity, or poor quality of the wording.*" (p. 6) **Evidence for task assignment → Acquire Resources / Develop Team (Resource Management):** "*The model of task assignment between performers uses the results of previous estimates to effectively assign tasks to individual team members.*" (p. 6) **Evidence for sprint planning → Develop Schedule (Schedule Management):** "*Intelligent planning of the IT project team's work… selecting tasks for sprinting.*" (p. 10–11)
RQ3: How are LLMs used to support SPM tasks?	LLMs are used to **analyze and improve the clarity of natural-language task descriptions**, which directly influences planning decisions and risk reduction.	"This model… **based on LLM, automatically analyzes the task description** to identify incompleteness…" (p. 6) "The method… **offers an edited version**, increasing the clarity of requirements." (p. 17)
RQ3.1: What mechanisms (e.g., automation, communication, information processing) are used by LLMs to affect SPM outcomes?	**Information Processing**, **Decision Support**, and **Automation** of text evaluation.	"The **recommendation model**… summarizes data… and generates final recommendations for the sprint plan." (p. 6) "The **model… automatically analyzes** the task description…" (p. 6)
RQ3.2: What are the outcomes affected by the mechanisms identified in RQ3.1?	Improving clarity of task descriptions leads to **reduced defect risk** and **preserved sprint value**.	"**Reducing the clarity of just two task descriptions increased the aggregated defect risk by 50% and decreased the integral sprint value by 10–15%.**" (Abstract, p. 1; repeated p. 18)
RQ4: What research design (e.g., empirical, case study, survey, interview) has been used to investigate the application of LLMs in SPM?	**Computational experiment** using generated task sets and **MILP optimization (OR-Tools)**.	"**Experimental evaluation** of the proposed model on benchmark datasets…" (p. 1) "**The optimization problem was solved by the MILP solver OR-Tools.**" (p. 11)
RQ5: How has the effectiveness of LLMs in SPM been measured?	Effectiveness was measured using **Sprint Value (V)**, **Risk (R)**, and the composite metric **W = V – R**.	"**Sprint value is an aggregate metric… Risks are defined…**" (p. 10) "**The criterion… W = V − R → max.**" (p. 10–11)

QA23:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**ChatGPT (OpenAI), Gemini Flash 1.5, Gemini Flash 2 (experimental), Claude (Anthropic), DeepSeek**	"**The same project description was given to four LLMs:** DeepSeek, ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google)." (p. 152) "ChatGPT, Gemini Flash 1.5, and Gemini Flash 2 consistently outperformed human participants…" (p. 154)
RQ2: What SPM tasks have been supported using LLMs?	**Agile Project Planning** **Subtasks:** **• Breaking work into structured tasks** **• Organizing tasks into sprints** **• Identifying risks and proposing mitigations** **• Ensuring adherence to Scrum planning principles**	"This study aims to evaluate the feasibility of LLMs in **Agile project planning**…" (Abstract, p. 150) "Project plans… focusing on key Scrum principles such as **task breakdown, sprint organization, and risk management**." (Abstract, p. 150)
RQ2.1: What PMBOK practices correspond to these tasks?	**Collect Requirements (Scope Management)**, **Define Scope (Scope Management)**, **Develop Schedule (Schedule Management)**, **Identify Risks (Risk Management)**, **Plan Risk Responses (Risk Management)**.	**Task breakdown → Collect Requirements & Define Scope (Scope Management):** "Project plans… included **task breakdown**." (Abstract, p. 150) **Sprint planning → Develop Schedule (Schedule Management):** "**Sprint organization** was evaluated…" (Abstract, p. 150) **Risk management → Identify Risks & Plan Risk Responses (Risk Management):** "**Risk management** was one of the evaluation criteria." (p. 153)
RQ3: How are LLMs used to support SPM tasks?	LLMs **generated complete Agile project plans** from prompts, including task structuring, sprint sequencing, and risk considerations.	"Each LLM was asked to **act as a Scrum Master and generate a project plan** using its understanding of Agile methodologies." (p. 152)
RQ3.1: What mechanisms (e.g., automation, communication, information processing) are used by LLMs to affect SPM outcomes?	**Automation** of project plan creation; **Information Processing** to structure tasks logically; **Decision Support** in selecting sprint sequences.	"LLMs demonstrated the ability to **produce project plans**…" (p. 154) "The quality of LLM-generated plans was **highly sensitive to prompt engineering**." (p. 154)
RQ3.2: What are the outcomes affected by the mechanisms identified in RQ3.1?	**Operational feasibility**, **task clarity**, **sprint organization quality**, and **adherence to Scrum principles**.	"ChatGPT, Gemini Flash 1.5, and Gemini Flash 2 **outperformed human Scrum Masters** in terms of **operational feasibility, task clarity, and sprint organization**." (p. 154)
RQ4: What research design (e.g., empirical, case study, survey, interview) has been used to investigate the application of LLMs in SPM?	**Comparative empirical study** using **expert evaluation of LLM-generated vs. human-generated project plans**.	"…**evaluated and ranked** based on the opinions of software development specialists." (p. 154)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Expert ranking** based on five predefined evaluation criteria (task clarity, sprint planning, risk management, adherence to Scrum principles, realism/feasibility).	Evaluation criteria listed on p. 153. Ranking results shown in Table II (p. 154).

QA27:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**GPT-4** and **GPT-3.5**, integrated via **LangChain**	"CogniSim is implemented in Python, using **OpenAI's GPT-4 and GPT-3.5 models**, and **LangChain** (LangChain, 2023) for LLM integration." (p. 388)
RQ2: What SPM tasks have been supported using LLMs?	**Project Lifecycle Coordination using role-based agent collaboration** **Subtasks:** **• Refining backlog items** **• Performing PI/Sprint planning** **• Delegating responsibilities across roles** **• Facilitating communication and coordination among team roles**	"Cognitive agents… **assist in backlog refinement, sprint planning, code implementation, and quality assurance**." (p. 387) "Cognitive agents powered by LLMs **collaborate within a simulated software environment… assuming key Agile roles in product management, architecture, development, and testing**." (Abstract, p. 385) "Agents… **manage project management, DevOps, QA, and development roles**." (p. 387)
RQ2.1: What PMBOK practices correspond to these tasks? *(Include PMBOK area in brackets)*	**Collect Requirements (Scope Management)**, **Define Scope (Scope Management)**, **Develop Schedule (Schedule Management)**, **Acquire Resources (Resource Management)**, **Develop Team (Resource Management)**, **Manage Communications (Communications Management)**.	Backlog refinement → **Collect Requirements / Define Scope (Scope Management):** "Product Owners **manage backlog prioritization**…" (p. 386) Sprint/PI planning → **Develop Schedule (Schedule Management):** "Agents collaborate with human stakeholders during **Program Increment (PI) preparation**…" (p. 386) Task delegation → **Acquire Resources / Develop Team (Resource Management):** "**Task delegation**… cognitive agents simulate roles such as Project Manager, DevOps Engineer, QA/Test Engineer…" (p. 387) Communication → **Manage Communications (Communications Management):** "Agents **improve communication** and coordination across teams." (p. 387)
RQ3: How are LLMs used to support SPM tasks?	LLMs are used to **simulate role-based decision-making**, **generate and refine user stories**, **analyze context**, and **automate communication and coordination** among agents.	"Each virtual agent… is **powered by an AI-driven LLM core**." (p. 388) "Cognitive agents… **automate routine tasks and optimize efficiency**." (p. 386)
RQ3.1: What mechanisms (e.g., automation, communication, information processing) are used by LLMs to affect SPM outcomes?	**Automation**, **Information Processing**, **Decision Support**, **Communication Support**.	"CogniSim integrates cognitive agents… to **emulate human-like decision-making, automate tasks, and streamline workflows**…" (p. 386) "Agents **interact and collaborate** to manage workflows." (p. 386–387)
RQ3.2: What are the outcomes affected by the mechanisms identified in RQ3.1?	**Improved decision quality**, **reduced manual workload**, **enhanced team coordination**, **increased efficiency and adaptability**.	"The framework… **improves decision-making**, increases efficiency, and **enhances collaboration**." (Abstract, p. 385) "Agents… **coordinate effectively, share information efficiently, and align objectives**…" (p. 389)
RQ4: What research design (e.g., empirical, case study, survey, interview) has been used to investigate the application of LLMs in SPM?	**Simulation-based evaluation** and **case study style demonstrations** within a **virtual environment**.	"A series of **simulations were conducted** to assess the effectiveness…" (p. 389) "Case studies… **demonstrate advancements** in task delegation, communication, and lifecycle management." (Abstract, p. 385)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Qualitative performance observations** during simulation, such as **communication quality**, **coordination effectiveness**, **decision-making quality**, and **alignment of objectives**. *(No quantitative metrics reported.)*	"Results showed that the cognitive agents **successfully replicated communication and decision-making processes**…" (p. 389) "Comparisons… indicated that agents **performed at a high level**, often matching or exceeding human-like analysis." (p. 389)

QA28:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**GPT-4**, specifically **GPT-4/GPT-4-Turbo/GPT-4o** as noted in different implementation steps.	"This paper explores the integration of Large Language Models (LLMs), **specifically GPT-4**, into software engineering…" (Abstract, p. 604) "**GPT-4-Turbo model** was selected…" (Model Setup, p. 608) "In this study, the **GPT-4o** model was selected due to its ability to handle large datasets." (Model Setup, p. 608)
RQ2: What SPM tasks have been supported using LLMs?	**Requirement Change Impact Analysis (RCIA)** **Subtasks:** **• Interpreting requirement change requests** **• Predicting multi-dimension impacts (technical / business / stakeholder)** **• Prioritizing impacts to support accept/reject decisions**	"This paper proposes… a methodology that leverages… **GPT-4 to analyze and predict the impacts of software requirement changes** before implementation." (Abstract, p. 604) "We then present the results… prioritizing the most critical impacts for decision-makers." (Ranked Results, p. 607) "Functional, business, technical, and **stakeholder impact** analysis." (Fig. 3 description, p. 608)
RQ2.1: What PMBOK practices correspond to these tasks?	**Perform Integrated Change Control (Integration Management)** — because the system evaluates and informs decisions on accepting/rejecting requirement changes. **Control Scope (Scope Management)** — because the impact analysis determines how requirement modifications affect system functionality. **Manage Stakeholder Engagement (Stakeholder Management)** — because the model evaluates **stakeholder impact** as part of change decisions.	**Perform Integrated Change Control:** "This study… assists requirements engineers to **make informed decisions regarding the acceptance or rejection of requirement change requests**." (Abstract, p. 604) **Control Scope:** "Impact analysis… examines how the change affects the system functionality… architecture… and business efficiency." (Implement & Testing, p. 608) **Manage Stakeholder Engagement:** "**Stakeholder impact analysis** examines these effects." (Fig. 3 explanation, p. 608)
RQ3: How are LLMs used to support SPM tasks?	GPT-4 performs **semantic analysis** of requirement documents and change requests, identifies affected components, and generates structured impact explanations.	"A large language model (LLM), such as **GPT-4, performs impact analysis**." (Impact Analysis via LLM, p. 607)
RQ3.1: What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing**, **Decision Support**, **Automation** of analysis.	"GPT-4… **interprets complex software documentation**…" (Abstract, p. 604) "Results… **prioritized for decision-makers**…" (Ranked Results, p. 607)
RQ3.2: What outcomes are affected by the mechanisms identified in RQ3.1?	**Improved accuracy**, **reduced manual effort**, **faster decision-making**, **more reliable impact prediction**.	"The findings… demonstrate that **GPT-4 can significantly improve the accuracy** of impact predictions…" (Conclusion, p. 609) "Expedite the analysis process far beyond that of traditional system analysts." (Discussion, p. 609)
RQ4: What research design has been used?	**Experimental evaluation** using **iTrust dataset** + **expert review validation**.	"We used the **iTrust dataset**…" (Data Collection, p. 608) "We conducted an **expert review**…" (Evaluation, p. 609)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Precision, Recall, F1-score** (confusion matrix) + **Expert scoring (Likert 1–10).**	Table III (Precision, Recall, F1-Score) (p. 608) "Expert assessment… average score of **7.6 out of 10**." (Evaluation, p. 609)

QA 30:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**BERT-base** and **BERT_SE** (fine-tuned transformer models). *(Note: These are encoder-based PLMs, **not** generative LLMs.)*	"We fine-tuned **BERT and BERT_SE** with a set of user stories and their respective functional size…" (Abstract, p. 604) "We trained two pre-trained language models (**BERT-base** [37] and **BERT_SE** [30])…" (p. 605–606)
RQ2: What SPM tasks have been supported using LLMs?	**Software Size Measurement (SSM) for effort and schedule estimation Subtasks:** • Measuring functional size of user stories (COSMIC CFP) • Automating size prediction using NLP • Supporting effort/schedule estimation decisions in organizations lacking expert workforce	"Software Size Measurement (SSM) plays an essential role in **effort and schedule estimation**…" (Abstract, p. 604) "We trained… models to **predict the functional size of software project requirements**…" (p. 605)
RQ2.1: What PMBOK practices correspond to these tasks?	**Estimate Activity Durations (Project Schedule Management)** — because functional size is used as the **primary input for schedule estimation**. **Estimate Costs (Project Cost Management)** — because size measurement informs **development effort and cost decisions**.	"Software Size Measurement… enables the acquisition of software size, which is the **primary input for development effort and schedule estimation**." (Abstract, p. 604) "Reliable SSM is therefore of critical importance for… **cost and schedule estimation**." (p. 604)
RQ3: How are LLMs used to support SPM tasks?	The models **predict COSMIC function sizes from natural-language user stories**, replacing manual size counting.	"We trained two NLP models… to **predict the functional size** of user stories." (p. 605)
RQ3.1: What mechanisms are used by LLMs to affect SPM outcomes?	**Information processing** (semantic understanding of requirements) and **automation** of size calculation.	"BERT models… provide **context-aware representations**…" (p. 604) "Thus, the model can be used for **quick size estimation without expert workforce**." (Conclusion, p. 609)
RQ3.2: What outcomes are affected by the mechanisms identified in RQ3.1?	• More consistent size estimates • Faster estimation • Reduced reliance on expert measurement • Ability to perform estimation with minimal resources	"This situation… becomes critical… therefore, organizations need to **perform objective SSM without an expert workforce**." (p. 604) "This model possesses an **acceptable margin of error for size measurement done quickly without requiring expertise**." (p. 607)
RQ4: What research design has been used?	**Exploratory case study** with dataset preparation, model training, and evaluation.	"In this research, we conducted an **exploratory case study**…" (p. 605)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**MAE**, **NMAE**, and **Accuracy** metrics using cross-validation.	"We used three metrics: **MAE, NMAE, and accuracy**." (p. 606) Prediction results in Table IV (p. 607)

QA 31:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**CodeBERT** (Transformer-based pre-trained bimodal model for code + NL). *(Note: CodeBERT is an encoder-based model, not a generative conversational LLM.)*	"We… investigate the applicability of using **CodeBERT** to predict the size and effort of software projects using the code as input." (Abstract, p. 603–604)
RQ2: What SPM tasks have been supported using LLMs?	**Software Size Measurement (SSM) and Software Effort Estimation (SEE)** from source code. **Subtasks:** • Predicting functional size using COSMIC data movement classification • Predicting Event-based size (interaction, communication, process) • Predicting development effort (person-hours) directly from code	"We introduce two exploratory case studies aimed at **predicting the functional size (COSMIC and Event-based size) and effort** of software projects from the code…" (Abstract, p. 603–604)
RQ2.1: What PMBOK practices correspond to these tasks?	**Estimate Activity Durations (Project Schedule Management)** — because functional size and effort are direct inputs to schedule estimation. **Estimate Costs (Project Cost Management)** — because effort prediction is used for cost/budget allocation.	"Software Size Measurement (SSM)… serves as the **primary input for development effort and schedule estimation**." (Introduction, p. 603)
RQ3: How are LLMs used to support SPM tasks?	CodeBERT is **fine-tuned to predict COSMIC functional size and Event-based size and effort directly from code**, replacing manual measurement.	"We train **CodeBERT** with a set of functions derived from the CodeSearchNet corpus… and with event-based size and effort values." (p. 604–605)
RQ3.1: What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing** (semantic code understanding) and **Automation** of size/effort estimation.	"CodeBERT… learns general-purpose representations to support downstream NL-PL applications." (Background, p. 604)
RQ3.2: What outcomes are affected by these mechanisms?	• Faster estimation • Reduced need for expert measurers • Feasibility of performing estimation with limited resources	"Organizations need to perform objective SSM and SEE **with minimal resources and without relying solely on expert workforce**." (Introduction, p. 603–604)
RQ4: What research design has been used?	**Two exploratory case studies** using real code datasets and controlled training/evaluation.	"In this research, we conduct **two exploratory case studies**…" (Abstract, p. 603)
RQ5: How has the effectiveness of LLMs in SPM been measured?	Classification accuracy (COSMIC) and regression metrics (MAE, NMAE, MMRE, PRED(30)).	Table 6 (Accuracy = 84.5%) and Tables 7–8 (MAE, NMAE, MMRE, PRED) (Results section, p. 606–607)

QA 32:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**GPT-3.5-Turbo** (zero-shot) **RoBERTa** and **SETFIT** (few-shot fine-tuned transformer models) *(Note: GPT-3.5 is the only **generative LLM** in the study; RoBERTa/SETFIT are transformer encoders.)*	"We compare the performance of the SETFIT classifier with the performance achieved by **GPT-3.5** in a **zero-shot** learning scenario." (p. 2) "We train and evaluate a model based on **SETFIT**… and compare its performance with **RoBERTa**." (p. 2)
RQ2: What SPM tasks have been supported using LLMs?	**Issue Report Classification to support Maintenance & Change Management Subtasks:** • Assigning issue labels (bug / feature / documentation / question) • Reducing label inconsistency across projects • Supporting prioritization and routing of issues to maintainers	"Issue-tracking systems are… essential tools… Effective labeling of issue reports is of paramount importance to support **prioritization and decision-making**." (p. 1) "We evaluate GPT-3.5-turbo… in a **zero-shot classification** scenario." (p. 2)
RQ2.1: What PMBOK practices correspond to these tasks?	**Control Scope (Project Scope Management)** — because correct issue classification determines whether a request becomes a **change to system scope** (e.g., bug vs enhancement).	"Issue reports manage requests for **changes**… Effective labeling… supports **prioritization and decision-making**." (p. 1)
RQ3: How are LLMs used to support SPM tasks?	GPT-3.5-Turbo is used in **zero-shot classification**, prompted with task description + label definitions.	"We compare… SETFIT… with **GPT-3.5 in a zero-shot learning scenario**, where the model is prompted…" (p. 2)
RQ3.1: What mechanisms are used?	**Information Processing** (interpret issue text) and **Automation** of label assignment.	"GPT-like models can **classify issue reports** without the need for fine-tuning." (p. 4)
RQ3.2: What outcomes are affected?	• Reduced manual labeling effort • More consistent labeling • Enables classification even without annotated datasets	"GPT-like models can achieve a performance comparable to the state-of-the-art **without the need for fine-tuning**… when a gold standard is not available." (p. 4–5)
RQ4: What research design has been used?	**Empirical evaluation** comparing SETFIT vs RoBERTa vs GPT-3.5 on manually verified dataset.	"We summarize the findings of two recent **empirical studies**…" (p. 2)
RQ5: How has effectiveness been measured?	**Precision, Recall, F1-score** per class and averaged.	"Evaluation… provided in terms of **precision, recall, and F1-measure**." (p. 3) Performance tables shown in Table 2 & Table 3 (p. 4).

QA 34:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**No generative LLM was used.** However, the study evaluates: • **fastText** (word embeddings + classifier) • **bi-LSTM** (neural network on embedded text) • **DistilBERT** (pre-trained transformer encoder) *(DistilBERT is **not** treated as an LLM in synthesis; it is an encoder model.)*	"We performed experiments using… fastText… LSTM… and **DistilBERT**." (Abstract, p. 316)
RQ2: What SPM tasks have been supported using LLMs?	**Effort Estimation in Agile Scrum (via Story Point Prediction) Subtasks:** • Predicting story points from user story text • Supporting sprint planning decisions • Helping teams determine workload and velocity	"Machine learning can play an essential role in **planning and estimating the project schedule**." (Abstract, p. 316) "This paper… solve[s] the problem of **predicting effort estimates in Agile Scrum**." (Abstract, p. 316)
RQ2.1: What PMBOK practices correspond to these tasks?	**Estimate Activity Durations (Project Schedule Management)** — because story points are directly used to estimate the duration (time) of user stories. **Estimate Costs (Project Cost Management)** — because duration estimations influence resource cost planning.	"Effective **effort estimation** in agile project planning is vital… helps… build product plans… and have better cost discipline." (Abstract, p. 316)
RQ3: How are LLMs used to support SPM tasks?	The models compute **text embeddings** from user story descriptions and classify them into **story point categories**, automating manual estimation.	"User-stories… contain a title [and] description… The models are trained **to predict the story-point**… " (Dataset section, p. 318–319)
RQ3.1: What mechanisms are used by models to affect SPM outcomes?	**Information Processing** (semantic analysis of requirements text) + **Automation** of estimation.	"fastText works due to the skipgram mechanism… LSTM… handles information effectively…" (Conclusion, p. 326–327)
RQ3.2: What outcomes are affected?	• Faster estimation • Reduced reliance on team expertise • More consistent effort estimates	"Our findings show… organizations can **benefit… and accurately predict project deadlines and schedules**." (Abstract, p. 316)
RQ4: What research design has been used?	**Comparative experimental evaluation** across multiple NLP and deep learning models on 90,000 user stories.	"The dataset consisted of **90,000 datapoints**… We performed experiments using… TF-IDF, fastText, LSTM, DistilBERT…" (Dataset, p. 319; Methods, p. 316–317)
RQ5: How has the effectiveness of the models been measured?	**Validation Accuracy** and **MAE (Mean Absolute Error)**.	"Table 2 shows… **validation accuracy**… Table 3 shows **MAE**…" (Results, p. 323–324)

QA 36:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**Open-source generative LLMs used: • Mistral 7B • Gemma2 9B • LLaMA 3.1 8B • Aya 8B • Qwen2 7.6B • Starling LM 7B**	"We evaluate… six widely used open-source **LLMs—Mistral, Gemma2, LLaMA3, Aya, Qwen2, and Starling**…" (p. 3)
RQ2: What SPM tasks have been supported using LLMs?	**Fine-Grained Bug Report Categorization (Issue Triage)** **Subtasks: •** Assigning bug reports into specific bug subcategories • Supporting prioritization and routing to responsible teams • Identifying misclassified or ambiguous bug reports	"Accurate classification of issues is essential for effective project management and timely responses." (p. 1) "We evaluate… **fine-grained bug report categorization**…" (p. 2)
RQ2.1: What PMBOK practices correspond to these tasks?	**Control Scope (Project Scope Management)** — because classification determines whether a report represents a change to system functionality. **Manage Stakeholder Expectations (Stakeholder Management)** — because routing determines who must act on the issue.	"Categorization… helps teams efficiently **prioritize and route issues**…" (p. 1)
RQ3: How are LLMs used to support the SPM task?	LLMs generate **bug category labels** using **prompt engineering strategies** (e.g., CoT, ToT, Plan-and-Solve, Multi-Agent Debate, Self-Consistency).	"We analyze 221,184 fine-grained bug report category labels generated by selected **LLMs using various prompt engineering strategies**…" (p. 1)
RQ3.1: What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing** (semantic interpretation of bug text) and **Decision Support** (providing category + rationale).	"Prompt engineering… influences output characteristics, control over outputs, and categorization performance." (p. 1)
RQ3.2: What outcomes are affected?	• Consistency of bug classification • Ability to detect misclassified bug reports • Improved dataset label reliability	"Examining label consistency… can **identify unclear reports and detect misclassifications** in human annotations." (p. 1–2)
RQ4: What research design has been used?	**Large-scale empirical study** evaluating prompt strategies × multiple LLMs × 1,024 bug reports.	"We analyze… six LLMs… using nine prompt types and four output configurations… for 1,024 bug reports." (p. 2)
RQ5: How has effectiveness been measured?	**Precision, Recall, F1-score, and Cohen's Kappa** (agreement between LLMs and human labels).	"We empirically quantify… performance measured by **precision, recall, and F1-scores**…" (p. 6) "We measure agreement using **Cohen's Kappa**." (p. 6)

QA 38:

RQ	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used?	ALBERT, BERT, CodeBERT, DeBERTa, DistilBERT, RoBERTa *(all fine-tuned transformer-based models)*	"We study 6 neural transformer-based language models… ALBERT, BERT, CodeBERT, DeBERTa, DistilBERT and RoBERTa." (p. 1013)
RQ2 – What Software Project Management (SPM) task(s) have been supported?	**Bug Triaging** including: • Assigning bug reports to developers • Assigning bug reports to system components	"The first step in managing bug reports is related to triaging a bug to the appropriate developer…" (p. 1012) "Assigning components or modules to bug reports is also not a trivial task." (p. 1012)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Acquire Resources** *(Project Resource Management)* — assigning issues to the developer best suited to work on them. **Monitor and Control Project Work** *(Project Integration Management)* — handling incoming defects and ensuring corrective action routing.	**Reasoned Mapping**: The task involves *allocating human resources* (developer assignment) and *managing ongoing project issue handling* (bug routing decisions).
RQ3 – How are LLMs used to support the task?	Fine-tuning transformer-based models to classify each bug report and assign it to the most likely **developer** or **component**.	"We fine-tune all transformer-based models for developer and component assignment." (p. 1015)
RQ3.1 – What mechanisms are used?	**Automation** of classification; **Information Processing** of bug text using contextual embeddings; **Model Comparison & Statistical Evaluation**.	"The Transformer architecture's self-attention mechanism can capture semantic information…" (p. 1013) "We conduct both quantitative and qualitative analyses…" (p. 1013)
RQ3.2 – What outcomes are affected?	Improved accuracy in triaging compared to older ML/DL approaches; DeBERTa performs best among transformer models; SVM surprisingly competitive in some developer assignment scenarios.	"DeBERTa performs significantly better than other transformer-based language models…" (p. 1016) "Somewhat surprisingly, the simpler TF-IDF-based SVM baseline performs best… on two of our four studied datasets." (p. 1016)
RQ4 – What research design has been used?	Empirical comparative study across **4 open-source project datasets** consisting of **136k bug reports**.	"Our study context consists of 136k bug reports… from four popular open source software repositories." (p. 1016)
RQ5 – How has effectiveness been measured?	**Developer assignment:** Top@K Accuracy (Top@1, Top@5, Top@10), Mean Reciprocal Rank (MRR). **Component assignment:** Precision, Recall, F1-score.	"We utilize Top@K Accuracy and MRR…" (p. 1015) "For component prediction, we utilize three metrics: Precision, Recall, and F1-score." (p. 1015)

QA 40:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**No generative LLM used.** Model used: **BERT-base-uncased** (pre-trained transformer encoder fine-tuned for classification).	"We describe a **BERT-based classification technique** to automatically label issues…" (Abstract, p. 1) "We used the **BERT pre-trained model** from Google AI." (p. 3)
RQ2: What SPM tasks have been supported using LLMs?	**Issue Type Classification (Issue Triage) in Software Maintenance Subtasks:** • Automatically labeling issue reports as *bug*, *enhancement*, or *question* • Reducing manual labeling effort in GitHub • Supporting maintainers in prioritizing and resolving issues	"Issue tracking… developers need to **triage new entries** to understand the nature of the report…" (p. 1) "We describe a BERT-based model to **predict an issue's type**." (p. 1)
RQ2.1: What PMBOK practices correspond to these tasks?	**Control Scope (Scope Management)** — determining whether an issue requests a *feature*, *bug fix*, or *question*, which affects **project scope decisions**.	"Labeling issues… supports **prioritizing issues/features to be developed**." (p. 1) → Prioritizing and determining scope of work.
RQ3: How are LLMs used to support the SPM task?	BERT is **fine-tuned** on labeled GitHub issue data and then used to **predict the issue type** of new reports.	"We trained and tuned a **multi-class classifier** to label GitHub issues automatically." (p. 2)
RQ3.1: What mechanisms are used?	**Information Processing** of issue descriptions using contextual embeddings; **Automation** of label assignment.	"We used the **bert-base-uncased** model and **fine-tuned** it for multi-class classification." (p. 3)
RQ3.2: What outcomes are affected?	• Reduced manual triage effort • Improved label consistency • Higher accuracy vs. FastText baseline	"Manual labeling… is error-prone and time-consuming." (p. 1) "Our approach **outperforms FastText** with an F1-score of 0.8586." (p. 2 & Table 2, p. 4)
RQ4: What research design has been used?	**Experimental evaluation** using a real-world labeled dataset of ~800k GitHub issues.	"The dataset contains **803,417 labeled issue reports**…" (p. 2)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Precision**, **Recall**, **F1-score** (micro-averaged due to class imbalance).	"We calculated **Precision, Recall, and F1-Score** for each class… micro-averaging…" (p. 4)

QA 41:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**No generative LLM is used.** Model used: **BERT (pre-trained transformer encoder)** used for semantic feature extraction.	"We propose a novel model… with the widely used pre-trained model, **BERT**." (p. 3) "BERT is used for semantic feature extraction…" (p. 6)
RQ2: What SPM tasks have been supported using LLMs?	**Software Effort Estimation** (main task) **Subtasks:** • Estimating effort for Allocated Requirements (ARs) in industrial projects • Using semantic textual representations of AR descriptions • Using expert features collected from team/project history • Supporting project planning decisions on required person-months	"Effort estimation is a critical task of software development planning and management…" (p. 1) "Our study aims to assist project managers in estimating efforts for long-term planning." (p. 3)
RQ2.1: What PMBOK practice(s) correspond to these tasks?	**Estimate Activity Durations** *(Project Schedule Management)* — because the task directly calculates effort/person-months needed for activities.	"Project managers are required to **estimate the number of person-months needed** to complete each AR…" (p. 2) → aligns with estimating duration/effort of planned activities.
RQ3: How are LLMs used to support the SPM task?	BERT is used to extract semantic embeddings from AR descriptions; these are integrated with expert features via a neural layer to predict effort values.	"We extract semantic features from software tasks using… **BERT**, and fine-tune it for effort estimation." (p. 6) "Fine-SE combines semantic features and expert features using a fully connected layer." (p. 7)
RQ3.1: What mechanisms are used?	**Information Processing + Prediction Automation** through integrated semantic + expert feature learning.	"Effort estimation model… integrates semantic and expert features." (p. 7)
RQ3.2: What outcomes are affected?	• Improved accuracy of effort estimation • Reduced subjective bias of expert-based estimation • Better performance vs. GPT-2 and Deep-SE baselines	"Fine-SE outperforms expert estimation by 32.0%–45.2%… and outperforms Deep-SE and GPT2SP by 8.9%-91.4%." (p. 1 and p. 8)
RQ4: What research design has been used?	**Model development and experimental evaluation** using industrial + open-source datasets.	"We compare… on 17 industrial projects and four OSS projects with more than 30,000 software tasks." (p. 1)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Mean Absolute Error (MAE)**, **Mean Magnitude of Relative Error (MMRE)**, **PRED(50)**.	"The measures… applied in this evaluation are **MAE, MMRE, and PRED(50)**." (p. 9)

QA 42:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1: What LLMs have been used to solve SPM tasks?	**No generative LLM used.** Models evaluated: • **RoBERTa-Large** (final selected model) • BERT • XLNet • DistilBERT • SBERT *(These are transformer encoder PLMs, not conversational LLMs.)*	"This paper introduces **ClassifAI**, an automated Issue Report Categorization approach built on the foundation of the **Transformer-based pre-trained RoBERTa-Large model**." (Abstract, p. 1) "We experimented with… **BERT, XLNet, DistilBERT, S-BERT and RoBERTa**…" (p. 2–3)
RQ2: What SPM tasks have been supported using LLMs?	**Issue Report Classification (Issue Triage) in Software Maintenance Subtasks:** • Classifying issue reports into *Bug*, *Enhancement*, or *Question* categories • Supporting routing of work to the correct responsible actor • Reducing manual effort and improving consistency in issue handling	"ClassifAI categorizes issue reports into **Bug report, Enhancement/feature request, and Question**…" (Abstract, p. 1) "Issue report classification… ensures streamlined and effective management." (Introduction, p. 1)
RQ2.1: What PMBOK practice(s) correspond to these tasks?	**Control Scope (Scope Management)** — because classification determines whether an issue proposes a *change* (enhancement) or a *defect fix*. **Manage Stakeholder Engagement (Stakeholder Management)** — routing issues ensures the correct stakeholder receives the request.	"Classification… ensures each issue report is **directed to the appropriate individual or team**…" (Introduction, p. 1)
RQ3: How are LLMs used to support the SPM task?	RoBERTa-Large is **fine-tuned** on labeled issue titles and descriptions to automatically predict issue type.	"We fine-tuned the **RoBERTa-Large** model on the refined dataset." (Abstract, p. 1)
RQ3.1: What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing** (contextual understanding of issue text) and **Automation** of categorization workflows.	"Transformer-based… models… classify natural language text effectively." (p. 1–2)
RQ3.2: What outcomes are affected by the mechanisms identified in RQ3.1?	• Reduced manual triage effort • Improved consistency in classification • Faster prioritization of issues	"Issue report classification can **effectively reduce the manual work**…" (Conclusion, p. 4)
RQ4: What research design has been used?	**Experimental evaluation** using the NLBSE'24 competition dataset of ~3000 labeled issue reports from 5 open-source projects.	"This dataset encompasses **3 thousand labeled issue reports** extracted from 5 real open-source projects." (Data Collection, p. 3)
RQ5: How has the effectiveness of LLMs in SPM been measured?	**Precision**, **Recall**, **F1-score**, and **Accuracy**.	"Evaluation metrics… include **precision, recall, and F1-score**." (Evaluation, p. 3) Performance values shown in Table 1 and Table 2 (p. 3–4).

QA 43:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What LLMs have been used to solve Software Project Management (SPM) tasks?	**GPT-3.5-Turbo family models:** • gpt-3.5-turbo-0301 • gpt-3.5-turbo-0613 • gpt-3.5-turbo-16k-0613	"The models adopted in this study are… **gpt-3.5-turbo-0301, gpt-3.5-turbo-0613, gpt-3.5-turbo-16k-0613**." (p. 6)
RQ2 – What Software Project Management tasks have been supported using LLMs?	**Issue Labeling / Issue Report Classification** within Software Maintenance. **Subtasks:** • Identify whether issue is Bug / Feature / Question / Documentation • Support maintainers in prioritizing and routing issues	"Issue labeling is a crucial task for the effective management of software projects." (Abstract, p. 1) "We investigate to what extent we can leverage GPT-like LLMs to **automate the issue labeling task**." (p. 1)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Control Scope (Project Scope Management)** — distinguishing bugs vs. new features corresponds to determining whether a change request alters scope. **Manage Stakeholder Engagement (Project Stakeholder Management)** — labeling supports correct assignment and communication among contributors.	"Labels… help maintainers **prioritize and coordinate work**." (p. 2)
RQ3 – How are LLMs used to support the task?	GPT-3.5 models are **prompted in zero-shot and few-shot modes** to classify issue title+body into categories.	"We explore a **zero-shot learning scenario**… We also investigate **few-shot learning**…" (p. 3)
RQ3.1 – What mechanisms are used?	**Information Processing** (understanding issue text) **Decision Support / Automation** (automatic label assignment)	Described in zero-shot prompt and classification pipeline (pp. 3–4)
RQ3.2 – What outcomes are affected?	• Reduced manual labeling effort • Comparable performance to fine-tuned BERT-based models • Supports dataset creation when labeled data is limited	"GPT-like models can **achieve a performance comparable** to state-of-the-art BERT-like LLMs **without the need for fine-tuning**." (p. 10) "Can be used to **reduce labeling costs**…" (p. 10–11)
RQ4 – What research design has been used?	**Experimental study** comparing GPT-3.5 zero-shot and few-shot prompting vs. fine-tuned SETFIT baseline, using manually verified GitHub issue dataset.	Methodology section (pp. 3–6)
RQ5 – How has effectiveness been measured?	**Precision**, **Recall**, **F1-score**, and **Cohen's Kappa** for agreement with human annotators.	Evaluation metrics table and discussion (pp. 7–10)

QA 49:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**PaLM (Pathways Language Model)** Additionally, **ChatGPT** was used to generate the synthetic dataset.	"This paper introduces the use of **PaLM** to facilitate automated Agile task creation and management…" (p. 6)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Agile Task Creation and Task Augmentation Subtasks:** • Generating structured task descriptions • Assigning responsible teams/roles • Predicting task priority and deadlines • Supporting project managers in workload planning	"PaLM… is used to **augment the information** obtained… A single query is augmented to include the task name, description, teams… and the priority." (p. 6) "The above fields are **auto-populated** based on the incoming request…" (p. 9)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Define Activities (Project Schedule Management)** — because tasks are being created and structured. **Estimate Activity Durations (Project Schedule Management)** — because task deadlines are predicted.	"The information obtained is used to **forecast potential deadlines** for the new ticket…" (p. 6)
RQ3 – How are LLMs used to support the SPM task?	PaLM generates task metadata fields (name, summary, priority, responsible teams) and combines them with enterprise data for scheduling.	"The PaLM LLM… augments the input ticket… to include task name, description, teams… and priority." (p. 6)
RQ3.1 – What mechanisms are used?	**Information Processing** of natural language input + **Decision Support** for task structuring and routing.	"Such augmentation… assists project managers by **infusing more context to incoming tickets**…" (p. 9)
RQ3.2 – What outcomes are affected?	• Increased consistency in task description and assignment • Improved visibility into planning • Reduced manual effort in task creation	"This paper attempts to **automate** Agile project management…" (p. 10)
RQ4 – What research design has been used?	**Methodology Proposal + Demonstration** using synthetic dataset and regression modeling.	"The dataset was **synthetically created**… validated by a project management professional…" (p. 8)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	No direct evaluation of PaLM output quality. Task duration prediction was evaluated using regression metrics: **MSE** and **R^2**.	"The mean square error (MSE) and R^2 score were used to determine the best performing regression models." (p. 7)

QA 53:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What LLMs have been used to solve Software Project Management tasks?	**GPT-3.5-turbo-1106** (fine-tuned to model personality traits and simulate team roles).	"In this paper, we select **'gpt-3.5-turbo-1106'** as the base model for our GenAI agent and conduct fine-tuning on this basis…" (p. 1649)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Team Composition & Role Simulation for Project Teams Subtasks:** • Analyzing team member personality traits • Identifying missing personality/role gaps • Introducing GenAI agents to simulate missing roles in project teams • Supporting collaboration and team diversity during project execution	"We propose a GenAI-empowered project management framework… integrating GenAI agents to **simulate team members playing different roles**." (Abstract, p. 1648) "The team analysis module classifies team members' personalities and roles, and **identifies gaps and personality diversity needs** within the team." (p. 1649)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Develop Team** *(Project Resource Management)* — because the LLM is used to enhance team composition and collaboration.	"This framework… aims to enhance team diversity by introducing GenAI agents modeled after a range of successful team member roles and personalities." (Abstract, p. 1648)
RQ3 – How are LLMs used to support the SPM task?	GPT-3.5-turbo is **fine-tuned on a personality dataset** and used to generate **virtual team member behavior**, enabling personality-driven team role simulation.	"Our GenAI agents are **fine-tuned** with a personality dataset… to simulate the traits of each team role and personality." (p. 1649)
RQ3.1 – What mechanisms are used?	**Information Processing** (interpret personality traits) **Simulation / Behavior Modeling** (generate role-aligned responses).	"Fine-tuning… equips ChatGPT with the cognition of personality traits." (p. 1649–1650)
RQ3.2 – What outcomes are affected?	• Increased team personality diversity • Enhanced collaboration potential • Improved ability to simulate missing team roles	"This framework… **cultivating diverse and successful project teams**…" (Abstract, p. 1648) "Fine-tuning… demonstrates significant improvements in the model's understanding… validating the feasibility of GenAI teammates." (p. 1651)
RQ4 – What research design has been used?	**Framework proposal** + **Experimental fine-tuning and evaluation** using personality datasets (FriendsPersona + Essay Dataset test).	"We trained the 'gpt-3.5-turbo-1106' model… using FriendsPersona… and tested on Essay dataset." (p. 1650–1651)
RQ5 – How has effectiveness been measured?	**Precision**, **Recall**, and **F1-Score** for personality classification accuracy.	Evaluation results table (p. 1651, Table III).

QA 56:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**No generative LLM used.** Models used for prediction: • **BERT-base** • **BERT_SE (domain-adapted BERT for Software Engineering)**	"We fine-tuned **BERT** and **BERT_SE** with a set of user stories and their respective functional size in COSMIC Function Points." (Abstract, p. 188)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Software Functional Size Measurement (FSM) & Early Effort Estimation Support Subtasks:** • Predicting COSMIC Function Point (CFP) size of user stories • Supporting early project scope and planning • Reducing dependency on FSM experts for estimation	"Software Size Measurement (SSM) plays an essential role in software project management as it enables the acquisition of software size, which is the primary input for development **effort and schedule estimation**." (Abstract, p. 188)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Estimate Activity Durations (Project Schedule Management)** — because predicted size is used to determine effort/time. **Estimate Costs (Project Cost Management)** — because functional size drives cost estimation.	"Size is the primary input for **effort and schedule estimation**…" (Abstract, p. 188)
RQ3 – How are LLMs used to support the SPM task?	User stories are encoded using BERT/BERT_SE embeddings, then a regression head predicts CFP size values.	"We trained two pre-trained language models… BERT-base and **BERT_SE**… for the **CFP prediction** task." (Method, p. 189–190)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing** of requirement text → **Numerical size prediction automation**.	"Functional size prediction… using deep learning-based NLP models." (Method, p. 189)
RQ3.2 – What outcomes are affected?	• Increases consistency of size estimates • Reduces need for expert measurement labor • Provides acceptable size accuracy for early planning	"Our results… achieved **74.4% accuracy with BERT_SE**… these results… demonstrate the practical utility of language models in SSM." (Abstract, p. 188)
RQ4 – What research design has been used?	**Exploratory Case Study** using real user stories from 16 open-source projects.	"In this research, we conducted an **exploratory case study**…" (p. 188–189)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**MAE (Mean Absolute Error) NMAE (Normalized Mean Absolute Error) Accuracy (rounded CFP values)**	"We used **MAE**, **NMAE**, and **accuracy**… to evaluate the success of the trained models." (Evaluation, p. 190)

QA 57:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**GPT-3.5-Turbo** (used for abstractive summarization of issue threads). Also used: **LexRank** (non-LLM) for extractive summarization.	"This paper… aims to automate the process of issue thread summarization using… **GPT-3.5-Turbo**, reducing the time spent…" (Abstract, p. 341) "We employ **GPT-3.5-Turbo**… for abstractive summaries." (Section III.A.2, p. 342–343)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Issue Thread Interpretation & Understanding (Software Maintenance Communication Support) Subtasks:** • Summarizing issue descriptions, discussions, and pull request links • Reducing time spent understanding historical issue context • Supporting developer comprehension before resolving issues	"Understanding issue threads is an essential aspect of software maintenance…" (Abstract, p. 341) "Developers… spend up to 50% of their time trying to understand issues…" (Introduction, p. 341–342)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Communications** *(Project Communications Management)* — because the goal is to improve clarity and comprehension among contributors.	"Our model's successful generation of brief, clear, and pertinent summaries… **boosts team communication**…" (Abstract, p. 341)
RQ3 – How are LLMs used to support the SPM task?	GPT-3.5-Turbo is prompted in **zero-shot mode** to generate abstractive summaries of entire issue threads (issue description + comments + pull request discussion).	"We have used **GPT-3.5-Turbo** to generate abstractive summaries… using the **zero-shot learning methodology**." (Abstract, p. 341; Section III.A.2, p. 343)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing + Abstractive Summarization** to condense lengthy collaborative conversation data into narrative summaries.	"Our approach taps into the potential of the **zero-shot learning methodology**, enabling the model to produce **context-specific summaries**…" (Abstract, p. 341)
RQ3.2 – What outcomes are affected?	• Reduced time to understand issues • Improved clarity in communication • Better support for onboarding and context recovery	"The successful generation of… pertinent summaries **boosts team communication and project management**…" (Abstract, p. 341)
RQ4 – What research design has been used?	**Experimental evaluation** using summarization + metric-based performance analysis on PI-Link dataset issue threads.	"We have utilized the **PI-Link dataset**… containing issue and pull request links…" (Section III.A.1, p. 342)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**ROUGE** (extractive summary evaluation) and **BARTScore** (abstractive summary evaluation).	"The performance… is assessed using **ROUGE and BART scores**…" (Abstract, p. 341) Section IV.A (p. 343–344).

QA 59:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**GPT-4-turbo** (used to generate user stories and test case specifications).	"We… developed a tool 'Geneus'… using **GPT-4-turbo** to automatically create user stories from software requirements documents." (Abstract, p. 791) "We applied **GPT-4-turbo**, the most intelligent LLM in the industry as of when the experiment was conducted." (Section VI, p. 799)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Requirements Engineering – Automated User Story Creation Subtasks:** • Extracting functional requirements from RE documents • Generating structured user stories (who / what / why) • Adding acceptance criteria and definition of done • Generating test case specifications linked to user stories	"We aimed to… develop a tool that takes the requirements documents from users and **delivers the detailed user stories**…" (Section I.A., p. 792) "Our methodology… generates user stories… followed by **test case specifications**." (Methodology, p. 794–795)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Collect Requirements (Project Scope Management)** — The LLM extracts and structures stakeholder needs into user stories. **Define Scope (Project Scope Management)** — Generated user stories clarify what functionality will be delivered.	"Understanding the software requirements and distilling them into individual unit tasks (user stories) is the crucial stage…" (Conclusion, p. 800)
RQ3 – How are LLMs used to support the SPM task?	GPT-4-turbo processes requirement documents using **Refine and Thought (RaT) prompting** to reduce noise and generate user stories and test cases in JSON.	"We propose a prompting strategy, **'Refine and Thought' (RaT)**… to improve the performance of the LLM in prompts with large and noisy contexts." (Abstract, p. 791) "Each step… is implemented using the **RaT prompting strategy**." (Methodology, p. 794–795)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Information Processing + Structured Output Generation** through multi-step prompting (RaT blocks).	"RaT prompting… refines… and then generates the user stories accordingly." (Section III, p. 794–795)
RQ3.2 – What outcomes are affected?	• Reduces manual time spent writing user stories • Improves clarity and completeness of user story specifications • Supports structured test case development	"This study aims to **reduce additional loads off the software engineers** and increase… productivity…" (Abstract, p. 791)
RQ4 – What research design has been used?	**Tool Design + Experimental Evaluation** using six real requirements documents.	"We collected six mid-sized RE documents…" (Section III.C., p. 795)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	• **BERTScore** (semantic similarity) • **AlignScore** (factual correctness) • **RUST Human Survey** (Readability, Understandability, Specificity, Technical-aspects)	"Along with manual evaluation using **RUST**…, automatic evaluation with **BERTScore** and **AlignScore**…" (Abstract, p. 791)

QA 60:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**GPT-4o** (integrated with knowledge graphs and RCA agents).	"By integrating the **GPT-4o model** with a knowledge graph for evidence retrieval… the model can identify deeper and deeper causes." (p. 17)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Root Cause Analysis (RCA) and Post-Incident Problem Prevention** in large-scale legacy IT systems. **Subtasks:** • Interpreting incident descriptions and logs • Asking iterative "Why?" chains to identify true root causes • Differentiating internal code issues vs. external/vendor causes • Detecting recurring patterns across historical incidents	"We use Generative AI agents combined with the **'Five Whys' technique**." (Abstract, p. 15) "Our approach prevents recurring failures by **identifying similar code patterns** in legacy systems, enabling proactive improvements." (Abstract, p. 15)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Control Quality (Project Quality Management)** — identifying and eliminating root causes of recurring defects. **Monitor and Control Project Work (Project Integration Management)** — using incident data to prevent future failures.	"This method uncovered that… issues… were due to **internal code problems**… enabling proactive improvements." (Abstract, p. 15)
RQ3 – How are LLMs used to support the SPM task?	LLMs act as **RCA agents**, iteratively applying "Why" reasoning, supported by evidence retrieved from an enterprise knowledge graph composed of SDLC artifacts.	"We developed a procedure to… construct a knowledge graph… supplying information to **analytical agents**." (p. 15–16) "**Five Whys Analysis Agent** automates the iterative questioning process." (p. 17)
RQ3.1 – What mechanisms are used?	**Reasoning + Evidence Retrieval** (LLM + Knowledge Graph) combined with **Iterative Root Cause Questioning** (Five Whys funnel).	"The agent then detects recurring patterns… and classifies the root cause." (p. 17)
RQ3.2 – What outcomes are affected?	• More root causes attributed to code defects rather than external blame • Shift from reactive fixes to proactive system improvement	"Over **70% of issues** previously attributed to external factors… were actually due to **internal code deficiencies**." (p. 15 and p. 18)
RQ4 – What research design has been used?	**Case Study** in a global financial institution analyzing **5,535 projects** and recurring incident records.	"In a case study, we analyzed **5,000 projects** and identified over **400 files** with the same root cause." (Abstract, p. 15)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	Incident management KPIs: • **Reduction in Major Incidents** (45%) • **Reduction in Change Failure Rate** (45.5%) • **Decrease in Lead Time to Deployment** (46.3%)	"Led to a **45% reduction** in major incidents… **45.5%** reduction in change failure rate… **46.3%** decrease in lead time." (p. 19)

QA 62:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**NR (No LLM used).** The paper uses **Sentence-BERT embeddings** for vector representation, not LLM-based reasoning or generation.	"Step 1: **BERT Embeddings.** To capture contextual semantics, **Sentence-BERT (SBERT)** generates high-dimensional vector representations…" (Section III.C.2, p. 127340)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Requirements Prioritization and Clustering (Backlog Refinement & Sprint Planning) Subtasks:** • Extracting and structuring requirements • Ranking requirements by business value and urgency • Modeling inter-requirement dependencies • Clustering requirements to support sprint planning	"This work proposes an AI-driven approach for **prioritizing the Product Backlog and optimizing Agile sprint planning**." (Section I, p. 127336)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Collect Requirements (Project Scope Management)** — extracting and structuring requirements.**Prioritize Work / Sequence Activities (Project Schedule Management)** — clustering requirements into prioritized sprint groups.	The task involves organizing requirements and prioritizing them for development planning (derived based on the SPM task; no direct PMBOK terminology in paper).
RQ3 – How are LLMs used to support the SPM task?	**Not applicable (No LLM used).** Instead, **SBERT embeddings + dependency graph + K-Means + PSO** are used.	"SAPC… leverages **BERT embeddings**, graph-based dependency modeling, and optimization techniques… for backlog prioritization." (Abstract, p. 127335)
RQ3.1 – What mechanisms are used?	**Semantic Embedding + Dependency Graph Modeling + Clustering Optimization**	"Requirement dependencies are captured by constructing a **directed graph** and computing requirement importance using **PageRank**." (Section III.C.2, p. 127341)
RQ3.2 – What outcomes are affected?	• More consistent and automated backlog ordering • Reduced manual prioritization effort • Data-driven sprint planning	"This approach **reduces manual effort**, enhances prioritization accuracy, and supports more **efficient sprint planning**." (Abstract, p. 127335)
RQ4 – What research design has been used?	**Experimental evaluation** using **12 real-world software projects** and performance metrics.	"The dataset comprises **functional requirements collected from 22 real-world software projects**, with 12 selected for testing." (Section III.B.1, p. 127338)
RQ5 – How has the effectiveness of LLMs been measured?	**No LLM effectiveness evaluated.** Instead: • **Silhouette Score**, **Davies-Bouldin Index** (for clustering) • **F1-score**, **MSE**, **Top@3 Accuracy** (for classification)	"The performance… was evaluated using **F1-score, Mean Squared Error (MSE), and Top@3 Accuracy**." (Section IV, p. 127345–127346)

QA 65:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**GPT-4** (used as a "virtual expert" to replicate expert decision-making in evaluating criteria and success factors).	"In this research we utilize the OpenAI model, which is **ChatGPT-4** due to its reasoning ability and cost effectiveness." (Model Selection, p. 1073)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Requirements Change Management (RCM) Prioritization in Global Software Development (GSD) Subtasks:** • Identifying critical success factors (CSFs) for managing requirement changes • Ranking these factors using decision-support • Comparing human expert and LLM-based prioritization outcomes to support governance decisions	"This study integrated Large Language Models (LLMs) with the **Fuzzy Best-Worst Method (FBWM)** to enhance prioritization accuracy and **decision support** in ARCM." (Abstract, p. 1071)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Monitor and Control Requirements Changes** *(Project Scope Management)* — because the task focuses on evaluating and prioritizing factors influencing change management decision-making.	The work centers on **prioritizing ARCM success factors** to guide change decision-making (Concept described throughout Section I & III).
RQ3 – How are LLMs used to support the SPM task?	GPT-4 is prompted using **role-based persona instructions** to act as a domain expert, selecting best/worst criteria and performing pairwise comparison judgments for FBWM.	"We utilized a prompt engineering technique to allow the **LLM to mimic a domain expert role**… The task has been decomposed into four main tasks… Best and Worst Criteria Selection… pairwise comparison." (Expert Input, p. 1073–1074)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Decision Support via Comparative Judgement** (LLM evaluates criteria like an expert) + **Natural Language Reasoning** through structured prompts.	"GPT-4 excels in managing diverse expert roles and in its ability to **justify its answers** by providing a clear rationale behind its selections." (p. 1074)
RQ3.2 – What outcomes are affected?	• Improved **consistency** of prioritization results • Reduced reliance on manual expert scoring • Demonstrated close similarity between LLM and human rankings	"The findings indicate that the **LLM-driven FBWM exhibit high reliability** in mirroring expert judgments…" (Abstract, p. 1071)
RQ4 – What research design has been used?	**Comparative Decision Evaluation Study** (Human experts vs. LLM-generated rankings, using FBWM).	"We compared and validated the prioritization outcomes derived from **human expert assessments** with those generated by **LLMs**." (Introduction, p. 1071)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	• **Consistency Ratio (CR)** in FBWM outputs • **MAE**, **RMSE**, **Spearman's ρ**, **Kendall's τ** comparing human vs. LLM ranking alignment	"We employed **MAE, RMSE, Spearman's Rank Correlation (ρ), and Kendall's Tau (τ)** to evaluate the similarity…" (Correlation Check, p. 1076–1077)

QA 66:

Research Question	Factual Answer	Exact Evidence (with page reference)
RQ1 – What Large Language Models (LLMs) have been used to solve Software Project Management tasks?	**LLaMA 3.1 (8B)** and **GPT-4** are integrated into the GDI system as interchangeable LLMs for generating queries and answers.	"We use Ollama to host a local instance of **Llama 3.1 (8B)**… and **OpenAI's GPT-4** into our system." (Implementation, p. 311)
RQ2 – What Software Project Management (SPM) tasks have been supported using LLMs?	**Work Item Navigation and Knowledge Retrieval in Issue Tracking Systems (ITSs) Subtasks:** • Answering developer onboarding questions about services, teams, and responsibilities • Assisting **trace link recovery** between work items • Supporting exploration of system relationships and dependencies	"We validate GDI through two proof-of-concept implementations… i. **onboarding** and ii. **trace link recovery** (TLR)." (Introduction, p. 308)
RQ2.1 – What PMBOK practice(s) correspond to these tasks?	**Manage Project Knowledge (Project Integration Management)** — because the system enables retrieving and reusing organizational knowledge to support developer tasks.	The system is explicitly used to support **knowledge access** and sharing in software teams. (Context described p. 308–309)
RQ3 – How are LLMs used to support the SPM task?	LLMs generate **Cypher graph queries** and **natural language explanations** based on the user's question and knowledge graph context, enabling guided exploration of work items.	"The GDI Core constructs a prompt… the **LLM generates a query**… Based on this prompt, the LLM generates a query… and then creates the **natural language answer**." (User Interaction, p. 309)
RQ3.1 – What mechanisms are used by LLMs to affect SPM outcomes?	**Retrieval-Augmented Generation (RAG)** extended with **knowledge graphs (GraphRAG)** to reduce hallucinations and improve traceability of how answers were derived.	"GraphRAG… incorporates knowledge graphs… providing supporting information to **understand how the LLM queried the graph**." (Abstract & Background, p. 308–309)
RQ3.2 – What outcomes are affected?	• Improved **access to project knowledge** • Better onboarding efficiency • More transparent and explainable system navigation • Reduced effort to locate work item relationships	"Our findings indicate that GDI is generally **perceived as user-friendly and accurate for onboarding** tasks… Participants expressed willingness to adopt the system." (Conclusion, p. 312)
RQ4 – What research design has been used?	**User-centered validation** involving **cognitive walkthroughs, interviews, and surveys** with practitioners in two companies.	"We conducted a **group cognitive walkthrough**… and **user-centered validation sessions** with practitioners." (Validation Method, p. 311)
RQ5 – How has the effectiveness of LLMs in SPM been measured?	**Perception-based evaluation metrics:** • Perceived usability • Perceived accuracy • Adequacy of supporting information • Perceived usefulness • Likelihood of adoption	"We evaluate GDI based on **five criteria**: perceived usability, perceived accuracy… perceived usefulness, and likelihood of adoption." (Research Questions, p. 309)