Vol. 4 No. 1 (2026)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

THE ROLE OF ARTIFICIAL INTELLIGENCE IN SOFTWARE-DEVELOPMENT PROJECT MANAGEMENT WITHIN INTERNATIONAL MEDIUM-SIZED ENTERPRISES (SMES): A SYSTEMATIC REVIEW

Rana Muhammad Usman Shahid[*1], Abdullah Amir[2], Furqan Hameed[3], Zarfia Gull[4], Imtiaz Hussain[5]

*[*1,2]Department of Computer Science, University of the West of Scotland, London Campus, UK*
[3]Department of Business, BPP University, London Campus, UK
[4]Hull University Business School, University of Hull, UK
[5]Department of Computer Science, University of Management and Technology, Sialkot Campus, Pakistan

[*1]rmusmanshahid@gmail.com,[2]abdullahamir615@gmail.com,
[3]engr.furqanhameed@gmail.com,[4]gzarfia@gmail.com,
[5]khan.imtiaz.ly@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

DOI: http://doi.org/10.5281/zenodo.21671417

Software-intensive small and medium-sized enterprises (SMEs) are adopting artificial intelligence (AI) to improve software-development project management (SDPM), from planning/forecasting and risk control to quality assurance/testing and developer productivity. To synthesize recent evidence (2023–2025) on (i) where AI is applied across the SDPM lifecycle in internationally active SMEs, (ii) the effects on delivery and quality outcomes, and (iii) governance practices that enable auditable, cross-border adoption.

We conducted a secondary-study systematic review (PRISMA-informed). Sources included peer-reviewed venues and authoritative grey literature (standards, regulatory texts, and programmatic guidance). Searches spanned academic databases and official portals for ISO/IEC 42001 (AI management systems), the NIST AI Risk Management Framework including the 2024 Generative AI Profile, PRINCE2 7, and the EU AI Act. Eligibility targeted SMEs (~10–250 employees) engaged in software development with international activity; interventions involved AI uses that influence SDPM; outcomes included DORA metrics (lead time, deployment frequency, change-fail rate, MTTR), schedule/cost variance, and defect/escape rates. From 1,100 records identified, 980 remained after deduplication; 110 full texts were assessed; 22 studies met inclusion criteria for qualitative synthesis.

Convergent evidence shows AI-augmented coding and testing deliver the most consistent near-term benefits (faster task completion, reduced defect leakage). Delivery performance improves when AI is layered onto strong engineering practices (continuous integration, automated testing, trunk-based development). Predictive planning and risk triage yield earlier variance signals in SMEs that implement minimum viable telemetry. Governance stacks built on ISO/IEC 42001, NIST AI RMF, and PRINCE2 7 support auditability, while the EU AI Act shapes provider/deployer obligations for international SMEs. Persistent barriers include skills and enablement gaps, data/telemetry immaturity, privacy/security risks (e.g., prompt leakage), and vendor lock-in.

AI can improve SDPM outcomes in internationally active SMEs when integrated with disciplined delivery and right-sized governance. We propose a three-layer framework—work level (co-pilots with review gates), flow level (DevOps analytics, test-impact and flakiness intelligence), and system level (AIMS/RMF/PRINCE2 governance)—plus a staged roadmap tied to measurable project metrics..

Keywords: *Artificial intelligence; software-development; project management; SMEs; Agile/DevOps; governance and risk; EU AI Act; ISO/IEC 42001; NIST AI RMF; PRINCE2 7.*

1. INTRODUCTION

1.1 Background and problem

Software-producing small and medium-sized enterprises (SMEs) operate with lean budgets, limited PMO capacity, and uneven engineering maturity, while competing in markets where customers expect short cycle times, high reliability, and audit-ready governance across borders. Since late 2022, advances in large language models (LLMs) and modern ML tooling have opened credible opportunities to accelerate coding, target testing more precisely, and improve forecasting and risk triage in software-development project management (SDPM). Syntheses and surveys in the software-engineering (SE) literature document a rapid expansion of LLM uses—from code completion and refactoring to requirements assistance and test generation—alongside recurring concerns about security, provenance, and evaluation rigor (e.g., benchmark validity and human-in-the-loop requirements) (Zhang et al., 2023; Hou et al., 2023).

At the delivery-systems level, longitudinal evidence from the DORA 2024 program suggests that AI primarily amplifies existing technical capabilities: teams already practicing continuous integration (CI), automated testing, and trunk-based development see clearer improvements in lead time, deployment frequency, and change-fail rate when they layer AI on top, whereas teams without these foundations benefit less (DORA, 2024; Valiulla, 2025). This pattern is consistent with experimental and field studies showing time savings and perceived quality gains from coding assistants, with the caveat that enablement and human review are essential to manage hallucinations and error propagation (e.g., controlled trials with professional developers; enterprise studies of Copilot adoption) (Wang, 2025; Gao, & Research, 2024; Kalliamvakou et al., 2024; Fan et al., 2024).

Testing is often the first adoption foothold for SMEs because value is tangible and feedback cycles are short. Systematic reviews and empirical evaluations published in 2023–2024 catalogue AI-powered tools for test-impact analysis, flakiness detection, failure clustering, and synthetic data generation; they report efficiency gains but advise incremental rollout under human oversight (Soha et al., 2024; Lin et al., 2024; related 2023 reviews in AI-for-testing). In planning and forecasting, recent secondary studies and industrial trials indicate that minimum viable telemetry—consistent defect/incident tagging, pipeline event logs, and a small set of throughput/quality signals—enables earlier variance detection and better estimate calibration, which are central to SDPM (Fan et al., 2024).

For internationally active SMEs, governance has become as important as technical effectiveness. The EU Artificial Intelligence Act (Regulation (EU) 2024/1689) establishes risk-tiered obligations and clarifies the roles of providers and deployers, with phased application beginning in 2024 (Act, 2024). Many internal SDPM tools will fall into "limited-risk" categories that emphasize transparency, but SMEs that provide AI affecting customers' processes—particularly in regulated domains—may face stricter documentation and conformity expectations. Complementary standards and guidance provide ready-made scaffolding at SME scale: ISO/IEC 42001:2023 defines a certifiable AI management system (AIMS) for policy, roles, risk management, monitoring, and improvement; and the NIST AI Risk Management Framework with its 2024 Generative AI Profile offers practical control language (e.g., data lineage, prompt-injection defenses, model change control, and continuous monitoring) (Benraouane, 2025; Hao, 2025; Gueorguiev, 2024; AI, 2024). Project-governance methods such as PRINCE2 7 now include AI-specific advice on where AI appears in plans and risk/quality artifacts, helping PMOs make AI-related decisions visible and auditable (Holubcová, n.d.).

Despite clear momentum, adoption in SMEs remains uneven. Comparative policy and enterprise surveys from 2024–2025 show growing intent to use AI but persistent capability gaps—especially in training and telemetry—leading to under-realized benefits (Prates, & Pereira, 2025; Brey, & van der Marel, 2024; OECD SME Digitalisation, 2024). Consistent with these findings, practitioner reports highlight that many SMEs attempt generative AI before implementing foundational digital practices (skills, data governance, and basic automation), which diminishes ROI and raises security and privacy risks (Das Chowdhury et al., 2024).

In short, the problem space facing internationally active SMEs is threefold: (1) identifying which AI uses measurably improve SDPM outcomes; (2) instituting right-sized governance that meets cross-border expectations without bureaucratic drag; and (3) overcoming capability bottlenecks in skills and telemetry so predictive methods and risk controls work as intended.

1.2 Objectives and research questions

This systematic review consolidates **2023–2025** evidence to address four questions:

1. Which AI applications measurably improve software-project outcomes in internationally active SMEs?
2. How do AI tools influence forecasting and risk/quality control in Agile/DevOps contexts?
3. Which governance and assurance practices enable auditable value at SME scale across borders?
4. What barriers and moderators (skills, data maturity, regulatory exposure) shape realized impact?

Our primary outcomes align with SDPM practice and contemporary delivery research: lead time, deployment frequency, change-fail rate, and mean time to restore (MTTR) (the DORA four), plus schedule/cost

variance and defect density/escape rate; governance indicators include documentation completeness and audit readiness.

1.3 Contribution and scope

This review offers a structured synthesis mapped to the SDPM lifecycle—requirements/estimation; planning/forecasting; risk/issue management; QA/testing; and flow performance/benefits realization—and complements it with a pragmatic governance layer that integrates ISO/IEC 42001 (certifiable AIMS), the NIST AI RMF + 2024 Generative AI Profile (LLM-era risk controls), and PRINCE2 7 (AI-aware project artifacts). Our population centers on SMEs (~10–250 employees) that operate across borders or supply EU customers, thus being in scope for the EU AI Act (Kilian et al., 2025; Act, 2024). Methodologically, we include peer-reviewed studies (systematic reviews, empirical/experimental papers and surveys) and authoritative standards/regulatory sources, triangulated with long-running delivery research (DORA) and high-quality industry reports to reflect the 2023–2025 state of practice (Singla et al., 2025; DORA, 2024). From 1,100 records identified, 980 remained after de-duplication, 110 full texts were assessed, and 22 studies met inclusion criteria for qualitative synthesis (see Figure 1, PRISMA).

2. Literature review

2.1 AI assistance for software engineering work

Across the 2023–2025 corpus, "evaluate-and-assist" adoption patterns dominate: LLM-based coding assistants for code completion and refactoring; AI help for documentation/summarization and knowledge retrieval; and targeted generation such as unit tests and test stubs (Zhang et al., 2023; Liu et al., 2024). Randomized and quasi-experimental studies with professional developers report measurable throughput gains and improved developer experience when assistants are deployed with enablement and review—e.g., enterprise-scale field experiments of GitHub Copilot and multi-site RCTs summarized by practitioner outlets (Shah et al., 2025; Wang, 2025; Gao, & Research, 2024; InfoQ, 2024). Although these studies were not SME-exclusive, the mechanisms—autocomplete for boilerplate, rapid navigation, context summarization—transfer to SMEs, where marginal hour-savings are more consequential (DORA, 2024). The same evidence stresses that benefits depend on disciplined code review, testing, and secure prompting; without this scaffolding, hallucinations, over-reliance, and weak prompt hygiene can erode quality or add debugging overhead (InfoQ, 2024). Broad surveys and mappings confirm the rapid expansion of LLM uses in SE while flagging persistent concerns about security, provenance, explainability, and evaluation rigor (Zhang et al., 2023; Liu et al., 2024). Mixed findings on downstream code quality—for example, reports of duplication and maintainability concerns in assistant-generated code—underscore the need for policy, static analysis, and reviewer checklists rather than blind reliance (Fu et al., 2025).

For internationally active SMEs, assistant tooling should be deployed with guardrails that respect data residency and confidentiality. The NIST Generative AI Profile (NIST AI 600-1) provides pragmatic control language—data origin/lineage, prompt-injection defenses, change management, and monitoring—that scales to SME contexts and yields auditor-friendly artifacts when combined with lightweight model cards and usage logs (NIST, 2024). These controls align with the broader NIST AI Risk Management Framework and are echoed in 2024 industry feedback to NIST (AI, 2024).

2.2 Predictive planning, estimation, and risk forecasting

Recent evidence describes two strands of AI-enabled planning in SDPM: (i) models trained on local histories (work-item and pipeline telemetry) and (ii) transfer approaches adapted to the SME's context (Ajorloo et al., 2024). Systematic and scoping reviews through 2024–2025 report improved estimate calibration and earlier variance signals when telemetry is adequate and models are monitored for drift. Even with modest data volume, standardizing defect/incident taxonomies and capturing pipeline events (build/test outcomes, change-fail patterns, code-review latency) unlocks predictive leverage. These findings converge with longitudinal delivery research: predictive analytics add value once *minimum viable telemetry* is in place.

At commit granularity, change-risk scoring has emerged as a practical mechanism to triage diffs for extra tests and reviews; publicly documented implementations show how code signals and metadata can forecast

incident-prone changes so PMs can gate risky diffs before release (Allam, 2025). In SME settings, this prioritization step maps directly to project risk logs and quality plans and can reduce MTTR and rework when combined with targeted testing.

2.3 QA/testing intelligence

Testing is one of the most tractable, high-ROI adoption areas for SMEs. Systematic reviews and empirical evaluations from 2023–2025 document growth in AI-assisted test selection (test-impact analysis), flaky-test detection, failure clustering, and synthetic test-data generation (Padmanabhan, 2024). Reported effects include shorter regression cycles, fewer escaped defects, and better focus on risky changes—provided a human remains in the loop (Lin et al., 2024). Cross-project studies also note that flaky-test predictors built on within-project features can degrade when moved across projects, so SMEs should prefer log-based features and local calibration. Position papers and theses in 2024–2025 analyze sources of flakiness in AI-enabled systems (e.g., nondeterminism), and early experiments in LLM-assisted flaky-test repair show promise for narrow cases but still require conventional techniques and manual oversight (Alaharju, 2024; Soha et al., 2024; Wadström er, 2025).

Security-focused analyses warn that plausible assistant-generated changes can embed subtle weaknesses; large-scale 2025 evaluations report elevated flaw rates in AI-generated code across languages, reinforcing the need for SAST/DAST alongside assistants (Veracode, 2025). For international SMEs expected to pass enterprise audits, demonstrating review evidence, coverage improvements, and change-risk rationale is critical.

2.4 Governance and regulation

For cross-border SMEs, three governance anchors recur. First, ISO/IEC 42001:2023 defines a certifiable AI management system (AIMS) covering policy, roles, risk management, monitoring, and continual improvement; guidance from vendors and assurance firms shows how SMEs can implement 42001 in a lightweight way and link it to existing ISO/IEC 27001 controls (Gueorguiev, 2024; KPMG, 2024; Reivo, 2024). Second, the NIST AI RMF and 2024 Generative AI Profile provide control language tailored to LLM-era risks—data lineage, prompt-injection defenses, model change control, and continuous monitoring—complementing an AIMS (AI, 2024). Third, PRINCE2 7 introduces AI-aware advice on where AI appears in plans, risk/issue logs, and quality artifacts so PMOs can make AI-related decisions visible and auditable.

Regulatorily, the EU Artificial Intelligence Act (Regulation (EU) 2024/1689) adds a risk-based framework with explicit provider and deployer obligations, phased from July 2024. Many internal SDPM supports (coding assistants, testing heuristics) will be limited-risk and require transparency/user information, whereas SMEs that provide AI in sensitive domains may face stricter documentation and, later, conformity assessment (European Parliament, 2025). Author summarize scope and timelines, which SMEs can operationalize by keeping model cards, intended-use statements, data-lineage records, and a lightweight AI risk register (Chen, & Deng, 2024; Turgonyi, 2025).

2.5 Adoption barriers

Two structural gaps recur across policy and practitioner surveys: skills/training and data/telemetry readiness. Many SMEs report insufficient training and uneven telemetry, which erodes the returns from assistants and testing intelligence (Zamfirescu-Pereira et al., 2025; Brey, & van der Marel, 2024). Baseline telemetry—consistent defect/incident taxonomies, unified pipeline logs, and agreed delivery/quality metrics—is often the tipping point for predictive use cases. Additional frictions include privacy/IP concerns (fear of prompt leakage), explainability expectations from enterprise customers, and vendor lock-in via proprietary telemetry or model artifacts; mitigations include enterprise-governed or self-hosted models, prompt scrubbing, model-use logging, and preferring open schemas with documented exit plans (KPMG, 2024).

3. Methodology

3.1 Design

We conducted a secondary-study systematic review guided by PRISMA 2020 and the PRISMA-S extension, adapting both to the heterogeneous and fast-moving 2023–2025 AI/software-engineering landscape (Page et al., 2021). Given the diversity of eligible evidence—randomized and quasi-experimental studies, field/case studies,

systematic reviews, and authoritative standards/regulatory texts—we specified a narrative synthesis with structured cross-tabulation rather than a meta-analysis, emphasizing convergent patterns over single-study effect sizes (Page et al., 2021). To balance recency and credibility, we triangulated peer-reviewed publications (e.g., SE surveys/SLRs and empirical evaluations), standards and regulatory sources (ISO/IEC 42001; NIST AI RMF with the 2024 Generative AI Profile; EU AI Act), and longitudinal industry research.

3.2 Eligibility criteria

Population/setting: SMEs (≈10–250 employees) engaged in software development and/or software-development project management (SDPM), with international exposure through cross-border operations or EU-market customers.

Interventions/phenomena: AI uses influencing SDPM, including planning/estimation/forecasting, risk and issue management, QA/testing, developer assistance, DevOps flow analytics, and governance/assurance mechanisms aligned to ISO/IEC 42001, NIST AI RMF/GAI, and PRINCE2 7 (Abiona et al., 2024).

Comparators: Any comparators were eligible; for qualitative and single-arm studies no explicit comparator was required (Page et al., 2021).

Outcomes: Delivery and quality outcomes aligned to contemporary SDPM practice—lead time, deployment frequency, change-fail rate, MTTR, schedule/cost variance, defect density/escape rate—and governance indicators such as documentation completeness and audit readiness.

Designs and time window: Systematic reviews/SLRs/MLRs, RCTs/experiments, field/case studies, standards/regulatory texts, and reputable surveys published 2023–2025 were eligible (DORA, 2024).

Exclusions: We excluded purely technical model benchmarks without SDPM outcomes, opinion pieces without methods, and pre-2023 materials except still-current standards or reporting guidance (Page et al., 2021).

3.3 Information sources and search strategy

We searched Google Scholar, ScienceDirect, SpringerLink, and arXiv for peer-reviewed work and current surveys, and consulted authoritative portals: DORA/Google Cloud (delivery research), NIST (AI RMF; NIST AI 600-1), ISO (ISO/IEC 42001), PeopleCert (PRINCE2 7), and EUR-Lex and specialist trackers for the EU AI Act (Rethlefsen et al., 2021). Search strings combined terms for *AI/GenAI/LLMs* with *software development/DevOps/testing*, *project management/PMO*, *SME/SMB*, and *international/EU AI Act*, and we performed forward/backward snowballing on keystone items (e.g., Peters, 2024; ISO/IEC 42001; NIST AI 600-1) (Kilian et al., 2025; Rethlefsen et al., 2021). The date limits were January 2023 to October 2025 inclusive to capture the LLM-era wave and the EU AI Act's publication and early implementation (Act, 2024).

3.4 Study selection, appraisal, and data extraction

Selection: Two reviewers independently screened titles/abstracts and then full texts using pre-specified criteria for SME relevance, SDPM outcomes, and recency, resolving disagreements by discussion (Page et al., 2021).

Appraisal: Because designs varied, we emphasized methodological clarity, outcome measurability, traceability of claims to data, and transferability to SME contexts, noting randomization where applicable and protocol transparency in reviews (Lin et al., 2024).

Extraction: We charted setting (SME/enterprise; international scope), AI use case, SDPM locus (planning/forecasting; QA/testing; developer assistance; governance), outcomes (e.g., DORA metrics, schedule variance, defect/escape), moderators (team size, engineering maturity, telemetry), and governance linkages (ISO/IEC 42001; NIST AI RMF/GAI; PRINCE2 7), plus EU AI Act roles/obligations.

PRISMA counts: The search identified 1,100 records (Jan 2023–Oct 2025). After de-duplication, 980 records progressed to title/abstract screening, of which 870 were excluded as out of scope. We assessed 110 full texts and excluded 88 (e.g., no SDPM outcomes, not SME, outside time window), leaving 22 studies in the qualitative synthesis (see Figure 1, PRISMA) (Page et al., 2021)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

3.5 Limitations

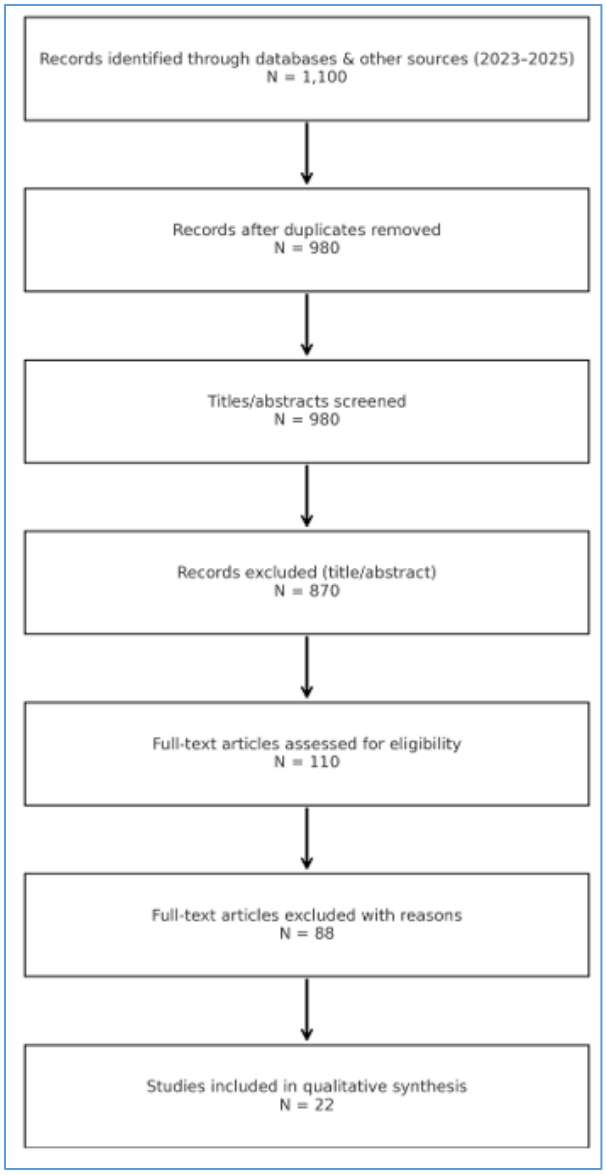

Figure 1 PRISMA-flow used in this review

We anticipated publication/novelty bias in 2023–2025 AI/SE work and therefore triangulated peer-reviewed studies with standards/regulatory anchors and longitudinal delivery research to stabilize conclusions. High heterogeneity in interventions and outcomes precluded meta-analysis; thus, we focused
on reproducible convergent patterns and explicit moderators (e.g., CI/testing maturity; telemetry quality). Finally, given rapid tool evolution, some findings may age quickly; we mitigated this by privileging governance frameworks (ISO/IEC 42001; NIST AI RMF/GAI; EU AI Act) that are relatively stable and by reporting the PRISMA counts transparently for replicability.

4. Results

4.1 Mapping AI interventions to SDPM outcomes

From the PRISMA process**,** 22 studies were included in the qualitative synthesis and are mapped to outcomes in Table 1; their themes are summarized in Figure 2. Across the 2023–2025 evidence base, five intervention clusters recur in internationally active software SMEs: (1) LLM-based coding assistants; (2) AI-assisted testing (test-impact analysis; flakiness detection; failure clustering; synthetic data); (3) predictive planning & risk triage; (4) flow analytics with AI; and (5) governance frameworks that make AI adoption auditable at SME scale. The most consistent signals of measured improvement appear where AI is layered onto robust engineering baselines—continuous integration (CI), automated testing, and trunk-based development—rather than used as a substitute (shorter lead times, lower change-fail rates, and better MTTR under disciplined delivery).

Coding assistance: Randomized and large-scale field studies report faster task throughput and improved developer sentiment when developers use tools like Copilot, with effects mediated by enablement and human-in-the-loop review (Wang, 2025; InfoQ, 2024). Although not SME-exclusive, the mechanisms—autocomplete, refactoring prompts, documentation synthesis—generalize to smaller teams that maintain review rigor and unit/integration testing. Still, mixed quality/trust signals (e.g., debugging overhead) reinforce the need for secure prompting, static analysis, and reviewer checklists in SME workflows (Noor, 2025).

AI-assisted testing: Systematic and multivocal reviews catalog a surge of tools since 2023 for test-impact analysis and flakiness detection; empirical evaluations show clear efficiency gains with practical limits, recommending incremental roll-outs with a human in the loop. Position papers and theses in 2024–2025 detail patterns in test instability (and the role of randomness) and show the value of log-based features for flakiness prediction, which travel better
across projects than purely code-centric models (Soha et al., 2024; Wadströmer, 2025).

Predictive planning & risk triage: Reviews and field reports indicate improved estimate calibration and
earlier variance detection where teams standardize telemetry (work-item histories, pipeline events) and monitor models for drift; effects are strongest when SMEs adopt minimum-viable telemetry (lead time, review latency, test pass/fail). Public documentation of change-risk scoring shows how commit metadata and code signals can prioritize additional testing or gated reviews ahead of release decisions—logic that maps neatly onto PMO risk registers and quality plans.

Flow analytics with AI: The 2024 DORA program—now a decade-long longitudinal series—reports that teams embedding AI within already disciplined delivery practices see the clearest gains across the four key metrics (lead time, deployment frequency, change-fail rate, MTTR). It also notes stronger outcomes where platform-engineering baselines are in place and product priorities remain stable (DORA, 2024).

Governance frameworks: Three anchors dominate for internationally active SMEs: ISO/IEC 42001:2023 (AIMS), the NIST AI RMF with the 2024 Generative AI Profile, and PRINCE2 7 with its AI guidance. ISO/IEC 42001 establishes policy, roles, risk controls, and continuous-improvement loops suitable for small organizations; the NIST GAI Profile contributes actionable controls for data lineage, prompt-injection, model change control, and monitoring; PRINCE2 7 provides project artifacts to make AI decisions visible in plans, risks, and quality approaches (Sanchez Salas, 2025; Gueorguiev, 2024; AI, 2024). Industry explainers show how these frameworks scale to SMEs and integrate with ISO/IEC 27001 (KPMG, 2024).

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Table 1. AI interventions mapped to SDPM outcomes (2023–2025)

AI intervention	SDPM locus	Outcomes	Representative recent sources
LLM-based coding assistants	Build; code review; backlog grooming	Faster task completion; higher developer satisfaction; quality maintained with human review/testing	Gao, & Research, (2024); InfoQ (2024); DORA (2024); Alaharju (2024)
AI-assisted test selection & flakiness detection	**QA/CI; change risk**	**Reduced regression time; lower escape defects; earlier detection of unstable tests**	**Lin et al., (2024); Soha et al. (2024); Wadströmer (2025).**
Predictive planning & risk triage	Estimation; forecasting; risk/issue logs	Better estimate calibration; earlier variance signals; focused mitigation	Ajorloo et al. (2024); DORA (2024).
Flow analytics with AI	DevOps; PMO dashboards	Shorter lead time; improved change-fail rate when paired with CI/testing	DORA (2024).
Governance frameworks (AIMS; RMF; PRINCE2 7)	PMO; compliance; assurance	Clear roles; audit trail; reduced regulatory exposure; repeatable decisions	ISO (2023); AI (2024); Holubcová, (n.d.); KPMG (2024); Microsoft (2025).

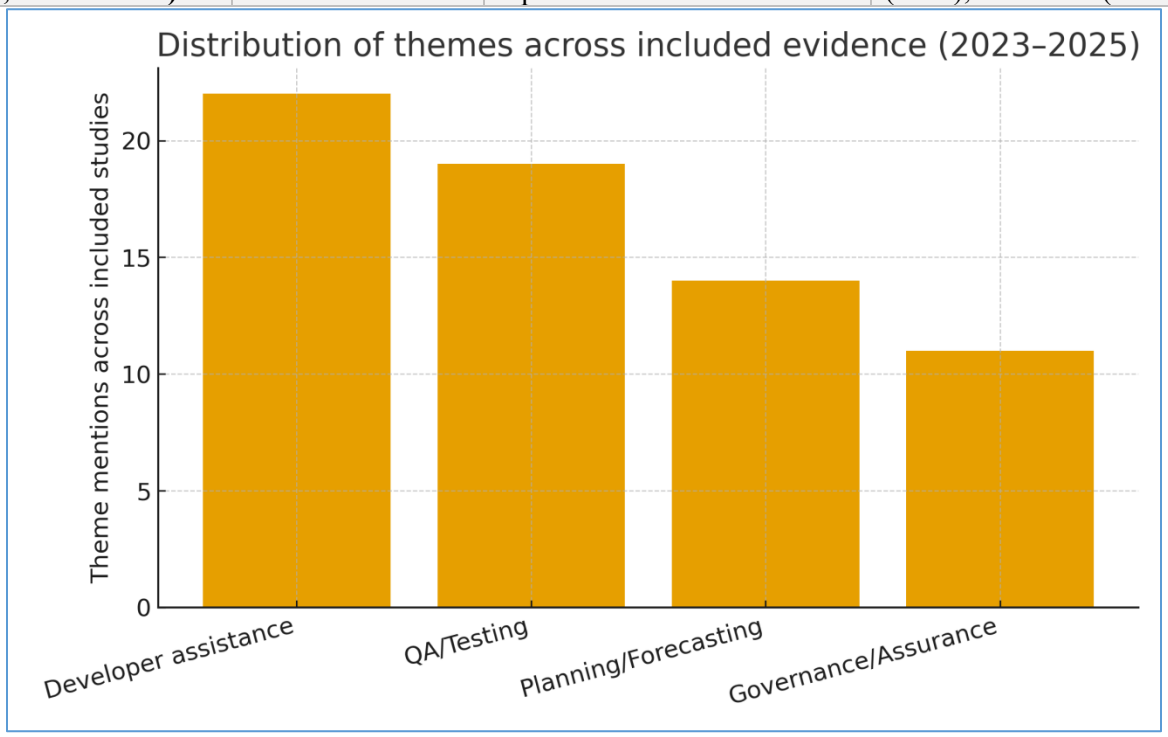

Figure 2 Distribution of themes across the 22 included studies (2023–2025)

4.2Thematic synthesis of outcomes

The 22 included studies concentrate most strongly on developer assistance and QA/testing, where interventions tie directly to everyday work and produce short feedback loops SMEs can manage. Convergent findings show assistants speed "plumbing" tasks and reduce cognitive load when paired with checklists and secure prompting; testing intelligence reliably trims regression cycles and reduces escape defects when teams sequence adoption (test-impact and flakiness first, generation later).

A smaller cluster addresses predictive planning and risk models. Studies show potential for earlier variance detection and more realistic schedules, but emphasize minimum-viable telemetry—consistent defect/incident categories, pipeline events, and PM-relevant measures such as lead time and review latency. Where implemented, change-risk scoring integrates smoothly into PM rituals by prioritizing high-risk commits for extra testing or gated reviews (Chen, & Deng, 2024).

Finally, governance emerges as an enabler rather than a brake when right-sized. ISO/IEC 42001 provides the scaffolding (policy, roles, risk); the NIST GAI Profile contributes concrete control language (prompt security, model change management); and PRINCE2 7 ties decisions to plan/risk/quality artifacts. SMEs using this trio satisfy large-customer and cross-border expectations without heavy bureaucracy and are better prepared for EU AI Act obligations (Kilian et al., 2025; Gueorguiev, 2024; AI, 2024).

4.3 Moderators and constraints

Engineering capability maturity: AI amplifies existing practices more than it replaces them. Improvements in lead time and change-fail rate are most visible when teams already run CI, maintain automated tests, and prefer trunk-based development (DORA, 2024).

Skills and enablement: 2024–2025 surveys highlight a gap between SME **intent** and **capability**: adoption is rising, but many SMEs under-invest in training; developer trust remains mixed, with many reporting debugging overhead on AI-generated code (Das Chowdhury, 2024; Noor, 2025).

Data readiness: Predictive models depend on consistent telemetry. SMEs often lack standardized defect/incident tagging and unified pipeline logs, limiting accuracy and drift detection. Testing evidence shows that adding log-based features and agreed taxonomies materially improves flakiness detection and failure clustering (Wadströmer, 2025).

Regulation and market access: International SMEs face EU AI Act obligations if they deploy or provide AI systems in the EU; risk tiers and the provider/deployer split shape documentation and transparency expectations. The Act entered the Official Journal in July 2024 with phased application, and official/curated portals summarize obligations and timelines (AI Act Tracker, 2024).

Market/worker sentiment: Developer adoption is high but trust lags; recent survey synthesis reports many developers use AI tools yet fewer than half fully trust outputs, citing buggy code and rework—reinforcing the need for quality gates, security scanning, and explicit prompts/policies.

5. Discussion

5.1 What works now for international SMEs

The synthesis points to two "now-ready" domains with repeatable value for internationally active SMEs: developer assistance and AI-aided QA/testing. Coding assistants reliably compress time spent on boilerplate, navigation, and documentation when paired with enablement and human review; field and randomized enterprise studies report higher task throughput and positive developer sentiment under these conditions. Testing intelligence—particularly test-impact analysis and flakiness detection—shortens regression cycles and reduces stealth defect escape, with strongest results when adoption is staged and a human remains in the loop (Lin et al., 2024). These wins are immediate because they fit existing sprint cadences and CI pipelines, and they can be tracked on delivery dashboards using DORA metrics such as PR cycle time, lead time, change-fail rate, and MTTR (DORA, 2024). For SMEs under cost and time pressure, the practical move is to wire assistants and test analytics directly into code review and CI scripting, making effects observable in routine ceremonies.

5.2 The next layer: predictive planning and risk triage

Second-phase opportunities arise once minimum-viable telemetry is in place (e.g., standardized defect/incident categories, pipeline event logging, basic delivery measures). Evidence from recent systematic reviews shows that even modest, well-labeled histories enable better effort estimation, earlier schedule-variance signals, and more focused mitigation. Commit-level change-risk scoring further helps PMs target extra tests and gated reviews before release decisions, translating to fewer customer-visible incidents and lower MTTR (DORA, 2024). Treat telemetry housekeeping as a prerequisite: agree common definitions, capture pipeline outcomes consistently, then iterate models under human oversight.

5.3 A right-sized governance stack

SMEs do not need enterprise bureaucracy to satisfy international expectations; they need credible, lightweight governance. Three components cover most needs:

1. **ISO/IEC 42001** (AIMS) to codify policy, roles (e.g., AI steward), risk registers, monitoring, and continual improvement (Gueorguiev, 2024).
2. **NIST AI RMF** plus the **2024 Generative AI Profile** to operationalize controls such as prompt security, model change control, data lineage, and monitoring (AI, 2024).
3. **PRINCE2 7** to surface AI decision points in plans, risks, quality approaches, and benefits reviews (Holubcová, n.d.).

Explain-and-assure notes from professional services and vendor compliance portals confirm that 42001 can be implemented incrementally and mapped to existing ISO/IEC 27001 controls—important for small teams that must satisfy enterprise procurement (KPMG, 2024). Adopting this trio early prevents retrofit costs and shortens security and procurement reviews for cross-border customers (KPMG, 2024).

5.4 International compliance posture

The EU AI Act (Regulation (EU) 2024/1689) imposes risk-tiered obligations and distinguishes providers from deployers; many internal SDPM tools will be limited-risk (transparency/user information), but SMEs that provide AI affecting customer processes—especially in sensitive domains—face stricter documentation and, over time, conformity assessment (Act, 2024). Trackers and official explainers outline phased implementation and documentation requirements; a lightweight compliance dossier—model cards, intended use, known limitations, data lineage, risk treatments—aligns with 42001 and the NIST generative-AI profile and eases due diligence (Turgonyi, 2025). Keep a single source of truth for AI systems (policy, role map, risk log, model cards, change log); it doubles as project evidence and external assurance (AI, 2024).

5.5 Integration framework and staged roadmap

A practical integration framework for SMEs comprises three layers:

- **Work layer (co-pilots + review gates):** Pilot LLM assistants in one or two squads; require secure prompting, code scanning, and human review. Baseline PR cycle time and tag AI-assisted commits to observe quality trends.
- **Flow layer (DevOps analytics + QA intelligence):** Add test-impact analysis and log-based flakiness detection; integrate change-risk scoring into release decisions; surface effects on DORA metrics in a shared dashboard (DORA, 2024; Peters, 2024).
- **System layer (AIMS + RMF + PRINCE2 7):** Institutionalize policy, roles, and artifacts (model cards, risk registers) and run quarterly reviews; prepare EU-AI-Act transparency notes for customer due diligence.

A **three-stage roadmap** follows:

Stage 1 (0–3 months): baseline DORA metrics; pilot assistants; institute prompt-hygiene and secure code scanning; reinforce review checklists (Gao, & Research, 2024).

Stage 2 (3–9 months): introduce test-impact analysis and flakiness detection; log pipeline events; draft ISO 42001-aligned policies; align with the NIST generative-AI profile (Krause, & Krause, 2025; AI, 2024).

Stage 3 (9–18 months): add predictive planning and change-risk scoring; run an AIMS internal audit; prepare EU-AI-Act transparency artifacts for sales/procurement.

5.6 Risks and mitigations

Hallucinations and quality: Maintain human-in-the-loop reviews; enforce unit/integration tests; adopt SAST/DAST; track defect origin for AI-assisted commits—practices that address observed trust gaps and debugging overhead in recent developer surveys (Alam et al., 2024; Noor, 2025).

Data leakage/IP: Prefer enterprise-governed or self-hosted models; scrub prompts; implement model-use logging and scoped access—controls aligned with the NIST Generative AI Profile (AI, 2024).

Skills gap: Fund targeted enablement (prompt hygiene, code-review playbooks, secure model use) and employ public SME programs; UK survey reporting shows capability shortfalls and demand for training (Prates, & Pereira, 2025).

Vendor lock-in: Favor open telemetry schemas and extractable model artifacts; document exit plans in the AIMS. Assurance and vendor explainers recommend mapping ISO 42001 to existing ISO 27001 to minimize duplication (Krause, & Krause, 2025; KPMG, 2024; Reivo, 2024).

6. Conclusion

For internationally active software SMEs, developer assistance and AI-aided testing offer the most reliable near-term gains and can be tracked via DORA metrics within normal delivery cadences. Predictive planning and risk triage become valuable once SMEs standardize minimal telemetry and adopt change-risk scoring. Rather than a compliance burden, a right-sized governance stack—ISO/IEC 42001 + NIST Generative AI Profile + PRINCE2 7—acts as an adoption scaffold and procurement advantage under the EU AI Act. The managerial takeaway is to build AI on disciplined engineering, invest in enablement and telemetry basics, and treat governance as a light but auditable system for visibility and reuse across projects.

REFERENCES

Abiona, O. O., Oladapo, O. J., Modupe, O. T., Oyeniran, O. C., Adewusi, A. O., & Komolafe, A. M. (2024). The emergence and importance of DevSecOps: Integrating and reviewing security practices within the DevOps pipeline. *World Journal of Advanced Engineering Technology and Sciences*, *11*(2), 127-133. https://www.researchgate.net/profile/Adebunmi-Adewusi/publication/379428586_The_emergence_and_importance_of_DevSecOps_Integrating_and_reviewing_security_practices_within_the_DevOps_pipeline/links/661d4f4c43f8df018d0e4011/The-emergence-and-importance-of-DevSecOps-Integrating-and-reviewing-security-practices-within-the-DevOps-pipeline.pdf

ACM Digital Library. (2024, Aug 8). *Flaky Tests in the AI Domain* (position paper). https://dl.acm.org/doi/10.1145/3643656.3643897

Act, A. I. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008,(EU) No 167/2013,(EU) No 168/2013,(EU) 2018/858,(EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU,(EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union L*, *1689*(12.7), 2024.

AI Act Tracker. (2024). *The Act texts*. https://artificialintelligenceact.eu/the-act/

AI lifecycle policy feedback (ITI). (2024, May 31). *Feedback to NIST AI-600-1*. https://www.itic.org/documents/artificial-intelligence/2024-05-31-ITIFeedbacktoNISTGenerativeAIProfile_Final.pdf

AI, N. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*. https://pathologyinnovationcc.org/s/R8b_1722118549750.pdf

Ajorloo, S., Jamarani, A., Kashfi, M., Kashani, M. H., & Najafizadeh, A. (2024). A systematic review of machine learning methods in software testing. *Applied Soft Computing*, *162*, 111805. https://www.academia.edu/download/115654648/ajorloo2024_.pdf

Alaharju, H. (2024). Ensuring performance and reliability in llm-based applications: A case study. https://aaltodoc.aalto.fi/server/api/core/bitstreams/6f2fc31f-7a20-4215-b7e2-4e27d7ca40f0/content

Alam, K., Mittal, K., Roy, B., & Roy, C. (2024). Developer Challenges on Large Language Models: A Study of Stack Overflow and OpenAI Developer Forum Posts. *arXiv preprint arXiv:2411.10873*. https://arxiv.org/pdf/2411.10873

Allam, H. (2025). Code Meets Intelligence: AI-Augmented CI/CD Systems for DevOps at Scale. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *6*(1), 137-146. https://ijaidsml.org/index.php/ijaidsml/article/view/181

Benraouane, S. A. (2024). *AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems*. Productivity Press. https://www.taylorfrancis.com/books/mono/10.4324/9781003463979/ai-management-system-certification-according-iso-iec-42001-standard-sid-ahmed-benraouane

Brey, B., & van der Marel, E. (2024). The role of human-capital in artificial intelligence adoption. *Economics Letters*, *244*, 111949. https://www.sciencedirect.com/science/article/pii/S0165176524004336

Chen, X., & Deng, Y. (2024). Evidential software risk assessment model on ordered frame of discernment. *Expert Systems with Applications*, *250*, 123786. https://www.sciencedirect.com/science/article/pii/S0957417424006523

Das Chowdhury, P., Renaud, K., & Rashid, A. (2024, September). When Data Breaches Happen, Where Does the Buck Stop?... and where should it stop?. In *Proceedings of the New Security Paradigms Workshop* (pp. 106-125). https://dl.acm.org/doi/pdf/10.1145/3703465.3703474

DORA. (2024). *Accelerate State of DevOps Report 2024*. https://dora.dev/research/2024/

European Parliament. (2025, February 19). *EU AI Act: First regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Fu, Y., Liang, P., Tahir, A., Li, Z., Shahin, M., Yu, J., & Chen, J. (2025). Security weaknesses of copilot-generated code in github projects: An empirical study. *ACM Transactions on Software Engineering and Methodology*, *34*(8), 1-34. https://arxiv.org/pdf/2310.02059

Gao, Y., & Research, G. C. (2024). Quantifying GitHub Copilot's impact in the enterprise with Accenture. *GitHub Blog*.

Gueorguiev, T. (2024, June). The Process Approach in Artificial Intelligence Management Systems. In *2024 9th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE)* (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/document/10600591

Hao, K. (2025). *Empire of AI: dreams and nightmares in Sam Altman's OpenAI*. Penguin Group. https://books.google.co.uk/books?hl=en&lr=&id=Oqo4EQAAQBAJ&oi=fnd&pg=PR4&dq=Altman+Solon.+(2024).+Enterprise+adoption+of+generative+AI:+Putting+Generative+AI+to+Work&ots=oifJY-OCtS&sig=tbqk_65o51CoCk6NWmkqhWwb-co&redir_esc=y#v=onepage&q&f=false

Holubcová, B. K. Project Management Tools For Implementing International Partnership Project. https://theses.cz/id/pjw2u6/Diploma_thesis_KH_final_Archive.pdf

InfoQ. (2024, Sept 24). *Study shows AI coding assistant improves developer productivity*. https://www.infoq.com/news/2024/09/copilot-developer-productivity/

Kilian, R., Jäck, L., & Ebel, D. (2025). European AI Standards–Technical Standardisation and Implementation Challenges under the EU AI Act. *European Journal of Risk Regulation*, 1-25. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E5157BA0391FFA9E1A3233E636005192/S1867299X25100329a.pdf/div-class-title-

european-ai-standards-technical-standardisation-and-implementation-challenges-under-the-eu-ai-act-div.pdf

KPMG. (2024). *ISO/IEC 42001: A new standard for AI governance*. https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html

Krause, D., & Krause, E. (2025). International Journal of Artificial Intelligence and Machine Learning. https://www.svedbergopen.com/files/1753935838_(4)_IJAIML2025145325NMEC7U_(p_43-50).pdf

Lin, S., Liu, R. Z. H., & Tahvildari, L. (2024). FlaKat: A Machine Learning-Based Categorization Framework for Flaky Tests. *arXiv preprint arXiv:2403.01003*. https://arxiv.org/pdf/2403.01003

Noor, N. (2025). Generative AI-assisted software development teams: opportunities, challenges, and best practices. https://lutpub.lut.fi/bitstream/handle/10024/169746/mastersthesis_Nouman_Noor.pdf?sequence=1

Peters, M. (2024). *Confident DevOps: The Essential Skills and Insights for DevOps Success* (Vol. 18). Kogan Page Publishers. http://repository.bitscollege.edu.et:8080/bitstream/handle/123456789/659/Confident%20DevOps%20(Mark%20Peters)%20(Z-Library).pdf?sequence=1

Prates, L., & Pereira, R. (2025). DevSecOps practices and tools. *International Journal of Information Security*, *24*(1), 11. https://link.springer.com/content/pdf/10.1007/s10207-024-00914-z.pdf

Reivo, J. (2024). Artificial intelligence governance, management and risk management-A look into EU AIA, standards and other frameworks from practical level. https://jyx.jyu.fi/bitstreams/b1bd0ec8-d76d-485d-9db6-727b42de4569/download

Ricca, F., García, B., Nass, M., & Harman, M. (2025). Next-Generation Software Testing: AI-Powered Test Automation. *IEEE Software*, *42*(4), 25-33. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11024091

Rossi, B. B., & Fontoura, L. M. AI-Based Approaches for Software Tasks Effort Estimation: A Systematic Review of Methods and Trends. https://www.scitepress.org/Papers/2025/132182/132182.pdf

Sanchez Salas, S. (2025). Digital Intimacy with AI Companions: When Vulnerability becomes a Business Risk-A Governance Framework for AI Companion Companies. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5435195

Shah, A., Chernova, A., Tomson, E., Porter, L., Griswold, W. G., & Soosai Raj, A. G. (2025, February). Students' Use of GitHub Copilot for Working with Large Code Bases. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1050-1056). https://dl.acm.org/doi/pdf/10.1145/3641554.3701800

Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. McKinsey & Company. (2025, Mar 12). *The state of AI (web)*. https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf

Soha, P. A., Vancsics, B., Gergely, T., & Beszédes, Á. (2024, April). Flaky Tests in the AI Domain. In *Proceedings of the 1st International Workshop on Flaky Tests* (pp. 20-21). https://dl.acm.org/doi/abs/10.1145/3643656.3643897

Turgonyi, D. A. (2025). Guardrails, Safeguards or Red Tape?: Competing Narratives of AI-Driven Disinformation Regulation in the European Union. https://www.diva-portal.org/smash/get/diva2:1969777/FULLTEXT01.pdf

Valiulla, A. (2025). An Empirical Study of Measurement Framework Adoption-DORA and SPACE: How Organizational Context Shapes Success and Failure. *Available at SSRN 5344643*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5344643

Visual Studio Magazine. (2024, May 16). *How good is GitHub Copilot?* https://visualstudiomagazine.com/articles/2024/05/16/copilot-benefits.aspx

Wadströmer, Z. (2025). Enhancing Flaky Test Detection with Log-Based Features: A machine learning approach. https://www.diva-portal.org/smash/get/diva2:1976980/FULLTEXT01.pdf

Wang, D. (2025). *In the future, software will be randomly generated* (Doctoral dissertation). https://repositories.lib.utexas.edu/server/api/core/bitstreams/7955a287-8e47-412e-9aa1-7b4b51315950/content

Zamfirescu-Pereira, J. D., Jun, E., Terry, M., Yang, Q., & Hartmann, B. (2025, April). Beyond code generation: Llm-supported exploration of the program design space. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). https://dl.acm.org/doi/pdf/10.1145/3706598.3714154