Quality Assessment — Grey Literature

Paper: Past the Demo: Modeling LLMOps Maturity for Production Adoption (A Rapid Multivocal Review)

Alonzo Nunez[1[0009-0001-3323-4319]], Christian Grévisse [2[0000-0002-9585-1160]] and Ma. Florencia Pollo Cattaneo [1[0000-0003-4197-3880]]

1 Method

Grey literature quality was assessed using a lean checklist adapted from Yasin et al. (2020), complemented by Garousi et al. (2019) multivocal review guidance. The assessment applies a pass/fail gate on three hard criteria, with inclusion requiring ≥ 2 of 3 passes.

1.1 Hard Criteria (determine inclusion)

Table 1. Hard Criteria

#	Criterion	Definition
C1	**Traceability**	Stable URL, identifiable authorship, date, and version control
C2	**Authority / Outlet**	Recognized vendor (tier-1), established research organization, university, or indexed outlet
C3	**Transparency / Reproducibility**	Methodology and criteria disclosed; operational artifacts, datasets, or verifiable assessment provided

1.2 Soft Signals (inform emphasis, do not determine inclusion)

Table 2. Soft Signals

#	Signal	Definition
S1	**Recency**	Publication or last update within 18 months of the search execution date (October 2025)
S2	**Objectivity**	Degree of independence from commercial interest; academic or editorially reviewed sources score higher than vendor-authored content promoting proprietary ecosystems
S3	**Novelty / Impact**	Introduces an original framework, method, dataset, or evaluation tool not substantially covered by other sources in the corpus

These signals were assessed per source and informed the relative qualitative emphasis given to each during synthesis. No formal quantitative weighting was applied.

2 Assessment Table

Table 2. Assessment Hard Criteria

ID	Source	Year	Type	C1	C2	C3	Score	Decision
G1	Microsoft Azure Blog	2024	Vendor blog	Y	Y	Y	3/3	**Include**
G2	AIM Research	2024	Industry report	Y	Y	P	2/3	**Include**
G3	AWS Prescriptive Guidance	2024	Vendor guide	Y	Y	Y	3/3	**Include**
G4	IBM Architectures	2024	Vendor docs	Y	Y	P	2/3	**Include**
G5	Özer — UEF Thesis	2025	Thesis/Report	Y	Y	Y	3/3	**Include**
G6	He et al. — arXiv	2024	Preprint	Y	Y	Y	3/3	**Include**
G7	Shi et al. — arXiv	2024	Preprint	Y	Y	N	2/3	**Include**
G8	Zhao et al. — arXiv	2025	Preprint	Y	Y	Y	3/3	**Include**
GQA5	DataStax Blog	2023	Vendor blog	N	Y	P	1/3	**Exclude**
GQA6	ZenML Blog	2024	Tech blog	Y	N	N	1/3	**Exclude**
GQA8	Yu & Ray — arXiv	2025	Preprint	Y	P	N	1/3	**Exclude** (RF2)
GQA13	KSV E-Journal	2025	Survey/Report	Y	P	N	1/3	**Exclude**

Legend: Y = Yes (pass) - P = Partial - N = No (fail) - **Score** = hard criteria passed out of 3

2.1 Soft Signal by Source

These signals informed qualitative emphasis during synthesis but did not determine inclusion or exclusion.

Table 3. Assessment Soft Criteria

ID	Source	Recency	Objectivity	Novelty
G1	Microsoft Azure Blog	Y	P	Y
G2	AIM Research	Y	P	P
G3	AWS Prescriptive Guidance	Y	Y	Y
G4	IBM Architectures	Y	P	P
G5	Özer — UEF Thesis	Y	Y	P
G6	He et al. — arXiv	Y	Y	Y
G7	Shi et al. — arXiv	Y	P	P
G8	Zhao et al. — arXiv	Y	Y	Y
GQA5	DataStax Blog	Y	P	P
GQA6	ZenML Blog	Y	N	P

ID	Source	Recency	Objectivity	Novelty
GQA8	Yu & Ray — arXiv	Y	P	P
GQA13	KSV E-Journal	Y	P	N

Legend: Y = Yes - P = Partial - N = No

3 Per-Source Rationale

3.1 Included Sources

G1 — Microsoft Azure Blog (3/3). Blog post with identified author, stable URL, and linked Microsoft Learn pages providing an operational assessment tool with scoring and level advancement guidance. Full transparency through published model, guide, and self-assessment. Soft signals support high emphasis in synthesis.

G2 — AIM Research (2/3). Landing page with six factors and five CMM-style levels from a recognized industry analyst. Transparency rated partial because the full whitepaper detail is gated behind registration; publicly available content is summarized. Soft signals support moderate emphasis in synthesis.

G3 — AWS Prescriptive Guidance (3/3). Official AWS documentation with four levels, six pillars, and explicit criteria per level. Available as both web and PDF. Stable URL with version control. Soft signals support high emphasis in synthesis.

G4 — IBM Architectures (2/3). IBM Architecture Center page describing a phased GenAI adoption model linked to reference architectures and patterns. Transparency rated partial because granular detail is dispersed across multiple IBM documentation pages. Soft signals support moderate emphasis in synthesis.

G5 — Özer, UEF Thesis (3/3). Master's thesis from the University of Eastern Finland applying a systematic methodology (PRISMA) to survey the LLMOps lifecycle, tools, and challenges. Full academic transparency. Soft signals support moderate emphasis in synthesis.

G6 — He et al. (3/3). arXiv preprint defining postconditions as a maturity measure for code LLMs with published benchmark code (PostcondGen) and reproducible evaluation protocol. Soft signals support high emphasis in synthesis.

G7 — Shi et al. (2/3). arXiv preprint applying LLMOps to personalized recommendation systems. University affiliations provide authority. Transparency rated as fail: narrative approach without reproducible protocol or published artifacts. Included at lower qualitative weight. Soft signals support low emphasis in synthesis.

G8 — Zhao et al. (3/3). arXiv preprint presenting a collaboration-of-experts framework for AIOps with an LLM classifier, RAG pipeline, and evaluation on the DevOps-Eval dataset. Method and artifacts fully disclosed. Soft signals support high emphasis in synthesis.

3.2 Excluded Sources

GQA5 — DataStax Blog (1/3). Vendor blog describing a four-arc maturity model. Traceability failed due to unstable URL following site migration. No reproducible evaluation protocol.

GQA6 — ZenML Blog (1/3). Comparative blog post synthesizing third-party maturity models. Authority failed as a vendor blog without editorial oversight. No original methodology or replicable artifacts.

GQA8 — Yu & Ray, arXiv (1/3). Preprint proposing a four-level maturity matrix for text-to-query along three axes. Excluded primarily under RF2 (out of scope for LLMOps adoption framing). Additionally, no validation artifacts provided.

GQA13 — KSV E-Journal (1/3). Narrative trends survey in a non-indexed institutional e-journal. Authority limited by outlet visibility; no systematic protocol or reproducible methodology.

4 References

- Yasin, A., Fatima, R., Wen, L., Afzal, W., Azhar, M., & Torkar, R. (2020). On using grey literature and Google Scholar in systematic literature reviews in software engineering. *IEEE Access*, 8, 36226–36243.

- Garousi, V., Felderer, M., & Mäntylä, M.V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, 106, 101–121.

- Kamei, F., et al. (2021). Grey literature in software engineering: A critical review. *Information and Software Technology*, 138, 106609.