D2.1 – Deliverable: State of the Art Report

LLM Evaluation for Systematic Literature Review (SLR) Data Extraction and Summarization

Project: AIDESL — ITEA Consortium
Work Package: WP2
Deliverable ID: D2.1
Status: Progress Report
Date: March 2026
Authors: Rushdi Shams, Staff AI Engineer, DistillerSR, Inc. rushdi.shams@distillersr.com

---

1. Introduction

This deliverable presents the current state-of-the-art investigation into Large Language Models (LLMs) under consideration for selection as the core AI backbone for Systematic Literature Review (SLR) data extraction and summarization tasks. The models investigated span two major families representing both proprietary API-based and open-source deployment paradigms.

The evaluation framework assesses each model across seven dimensions critical to SLR pipeline integration: feature representation, performance quality, scalability, efficiency, integration capabilities, open-source accessibility, and cross-domain performance.

This report constitutes an interim progress submission to the ITEA consortium, reflecting findings from the initial model benchmarking phase. Final model selection will be formalized in D2.2.

---

2. Legacy System Context

During the development and evaluation of the preceding system iteration, three open source models—OM1, OM2, and OM3 were assessed and deployed. OM1 was the most extensively validated of the three, serving as the production model for approximately one year; however, its overall performance across extraction accuracy, cross-domain generalization, and summarization quality ranged from medium to poor, ultimately rendering it insufficient for the demands of a reliable SLR pipeline.OM2 demonstrated stronger reasoning capacity but proved

prohibitively expensive to operate at the throughput volumes required, making it impractical for sustained deployment.OM3 offered fast inference speeds but exhibited unacceptable inaccuracy rates on structured extraction tasks. The limitations identified across all three models directly motivated the current investigation into the more capable and cost-balanced model set described in this deliverable.

---

3. Models Under Investigation

Five models have been shortlisted based on a preliminary landscape scan of available LLMs with demonstrated capabilities in scientific text processing:

#	Model	Parameters (approx.)	Access Type
1	CM1	Undisclosed	API (Proprietary)
2	CM2	Undisclosed	API (Proprietary)
3	CM3	Undisclosed	API (Proprietary)
4	OM4	17B active (109B total MoE)	Open Source (Apache 2.0)
5	OM5	17B active (400B total MoE)	Open Source (Apache 2.0)

---

3. Evaluation Framework

Each model is assessed against the following dimensions, chosen for their relevance to SLR automation workflows:

- **Performance**: Quality of extraction, summarization accuracy, and fidelity to source material in complex scientific literature tasks.
- **Response Speed**: Inference latency under realistic batch conditions (document-scale inputs).
- **Cost**: Operational cost per token/request at production throughput (API pricing or self-hosted infrastructure cost).
- **Open-Source Accessibility**: Availability of model weights, licensing permissiveness, and ease of local deployment.

- **Cross-Domain Performance**: Generalization ability across heterogeneous research domains (biomedical, engineering, social sciences, etc.).
- **Scalability**: Ability to handle increasing document volumes and context lengths without degradation.
- **Integration Capabilities**: Compatibility with existing SLR toolchains, API availability, and ease of pipeline embedding.

---

4. Model Evaluation Rubric

4.1 Summary Comparison Table

Dimension	CM1	CM2	CM3	OM5	OM4
Performance	🟢 High	🟢 High	🟡 Medium	🔴 Low	🟡 Medium
Response Speed	🟡 Medium	🟡 Medium	🟢 High	🟢 High	🟡 Medium
Cost	🟡 Medium	🔴 High	🟢 Low	🟢 Low	🟢 Low
Open-Source Accessibility	🔴 Closed	🔴 Closed	🔴 Closed	🟢 Open (Apache 2.0)	🟢 Open (Apache 2.0)
Cross-Domain Performance	🟢 High	🟢 High	🟡 Medium	🔴 Low	🟡 Medium
Scalability	🟢 High	🟢 High	🟡 Medium	🟡 Medium	🟡 Medium
Integration Capabilities	🟢 High	🟢 High	🟢 High	🟡 Medium	🟡 Medium

Legend: 🟢 High / Favourable — 🟡 Medium / Adequate — 🔴 Low / Limiting

---

4.2 Individual Model Profiles

Model 1: CM1

Overview: CM1 is mid-tier flagship model of its generation. It offers a strong balance of instruction-following accuracy and reasoning depth, positioning it as a high-capability option for

nuanced SLR tasks such as PICO extraction, study quality appraisal, and multi-document synthesis.

Dimension	Rating	Notes
Performance	🟢 High	Demonstrates strong understanding of methodological language, eligibility criteria, and structured extraction templates.
Response Speed	🟡 Medium	Suitable for asynchronous pipelines; not optimized for real-time interactive workloads.
Cost	🟡 Medium	API pricing is moderate; viable at medium throughput with careful prompt engineering to reduce token usage.
Open-Source Accessibility	🔴 Closed	Weights are not publicly available; access exclusively via API. Dependency on third-party availability and terms of service.
Cross-Domain Performance	🟢 High	Evaluated across biomedical, engineering, and social science corpora with consistently high relevance scoring.
Scalability	🟢 High	Supports long-context processing (200K token context window); suitable for full-article ingestion.
Integration Capabilities	🟢 High	RESTful API with robust SDK support (Python, TypeScript). Compatible with LangChain, LlamaIndex, and custom SLR toolchains.

Pros: Excellent instruction adherence; superior reasoning for ambiguous inclusion/exclusion criteria; extensive context window enables full-paper processing; strong tool-use and structured output capabilities.

Cons: Fully proprietary; operational dependency on API availability; higher cost at scale compared to open-source alternatives; no local deployment option for data-sensitive projects.

SLR Suitability: High. Recommended for tasks requiring deep comprehension of scientific methodology, complex eligibility adjudication, and multi-step summarization chains.

Model 2: CM2

Overview: CM2 is an updated release within the family of models it belongs to, offering incremental capability improvements over Sonnet 4 with enhanced agentic performance and

extended reasoning capacity. It represents the highest-capability option under evaluation, at a premium cost.

Dimension	Rating	Notes
Performance	🟢 High	Marginally stronger than Sonnet 4 on complex multi-step extraction and hierarchical summarization tasks.
Response Speed	🟡 Medium	Comparable to Sonnet 4; extended thinking features may introduce additional latency in reasoning-intensive tasks.
Cost	🔴 High	The highest-cost model in the evaluation set; may present budget constraints at scale.
Open-Source Accessibility	🔴 Closed	Proprietary API only; same constraints as Sonnet 4.
Cross-Domain Performance	🟢 High	Demonstrated strong generalization across heterogeneous scientific domains in initial benchmarks.
Scalability	🟢 High	Extended context support; improved handling of long-form agentic workflows.
Integration Capabilities	🟢 High	Full API compatibility with same toolchain integrations as Sonnet 4.

Pros: Best-in-class performance for sophisticated multi-document reasoning; enhanced agentic capabilities beneficial for automated SLR workflow orchestration; improved handling of ambiguous scientific language.

Cons: Cost is a significant constraint for large-scale SLR corpora; fully closed-source; incremental performance gains over Sonnet 4 may not justify the cost differential for all SLR task types.

SLR Suitability: High (with cost caveat). Optimal for high-complexity tasks where extraction quality is paramount and throughput volume is manageable within budget constraints.

Model 3: CM3

Overview: CM3 is lightweight, speed-optimized model designed for high-throughput, lower-complexity tasks. Within an SLR pipeline, it is best suited for screening and classification stages rather than deep data extraction.

Dimension	Rating	Notes
Performance	🟡 Medium	Adequate for title/abstract screening and binary inclusion decisions; less reliable for nuanced extraction requiring domain depth.
Response Speed	🟢 High	Fastest API response times among the family of models evaluated; well-suited for high-volume screening.
Cost	🟢 Low	Lowest cost among models; highly economical at scale.
Open-Source Accessibility	🔴 Closed	Proprietary API; same accessibility constraints as other models from the family.
Cross-Domain Performance	🟡 Medium	Acceptable generalization for surface-level classification tasks; reduced reliability for deep cross-domain synthesis.
Scalability	🟡 Medium	Efficient at scale for lightweight tasks; context window smaller relative to Sonnet variants.
Integration Capabilities	🟢 High	Full API compatibility; well-documented for pipeline integration.

Pros: Extremely fast inference enables rapid screening of large literature corpora; low cost enables deployment at full SLR scale; effective for structured classification tasks with well-defined criteria.

Cons: Reduced capability for complex extraction tasks; not recommended as a sole model for full-text data extraction; limited depth in cross-domain scientific reasoning.

SLR Suitability: Medium. Recommended as a complementary first-pass screening model in a multi-tier pipeline (e.g., Haiku for screening, Sonnet for full-text extraction), not as a standalone solution.

Model 4: OM5

Overview: OM5 is a Mixture-of-Experts (MoE) model with 17 billion active parameters from a total of approximately 400 billion, released by under the Apache 2.0 licence. While architecturally innovative, initial evaluations indicate performance limitations relative to frontier proprietary models on complex scientific text tasks.

Dimension	Rating	Notes

Performance	🔴 Low	Initial benchmarks show reduced reliability on complex SLR extraction tasks; struggles with nuanced eligibility adjudication and structured scientific output.
Response Speed	🟢 High	MoE architecture activates only a subset of parameters per token, resulting in fast inference at deployment.
Cost	🟢 Low	Open-source weights enable self-hosted deployment with infrastructure costs only; zero per-token API fees.
Open-Source Accessibility	🟢 Open	Fully open under Apache 2.0 licence; model weights freely available for research and commercial deployment.
Cross-Domain Performance	🔴 Low	Demonstrates limited generalization across diverse scientific domains in preliminary assessments; performance drops noticeably outside training distribution.
Scalability	🟡 Medium	Self-hosted deployment requires substantial GPU infrastructure; scalability is contingent on available compute resources.
Integration Capabilities	🟡 Medium	Compatible with HuggingFace Transformers, vLLM, and Ollama; integration requires more engineering effort than fully managed API solutions.

Pros: Full data sovereignty via local deployment; zero licensing fees; MoE architecture offers efficient inference; open weights enable fine-tuning on domain-specific SLR corpora.

Cons: Performance significantly below frontier models on complex tasks; cross-domain generalization limitations may introduce extraction errors in heterogeneous SLR corpora; self-hosting introduces operational overhead and hardware requirements.

SLR Suitability: Low for current off-the-shelf deployment. Potentially viable as a fine-tuning candidate if domain-adapted on SLR-specific training data, which remains a planned activity in subsequent work packages.

Model 5: OM4

Overview: OM4 is a more efficient model variant family of models it belongs to, with 17 billion active parameters from a 109 billion total MoE architecture. It offers a more balanced trade-off than Maverick between inference speed and task complexity, while retaining the open-source accessibility of the family of the model.

Dimension	Rating	Notes
Performance	🟡 Medium	Outperforms Maverick on structured extraction tasks in initial testing; does not reach the reliability threshold of variants of the model family for complex multi-step reasoning.
Response Speed	🟡 Medium	Moderate inference speed; slower than Maverick due to denser architecture but comparable to models of the same famly at equivalent hardware.
Cost	🟢 Low	Open-source; self-hosted infrastructure cost only.
Open-Source Accessibility	🟢 Open	Apache 2.0; fully open weights available for local deployment and fine-tuning.
Cross-Domain Performance	🟡 Medium	Adequate generalization across core scientific domains; performance more consistent than Maverick across domain shifts.
Scalability	🟡 Medium	Requires multi-GPU setup for production throughput; manageable with standard research compute infrastructure.
Integration Capabilities	🟡 Medium	Compatible with major open-source inference frameworks; integration feasible with moderate engineering investment.

Pros: Best performance among open-source candidates; data sovereignty through local deployment; fine-tuning potential on SLR-specific data; Apache 2.0 licensing supports all planned use cases.

Cons: Performance gap relative to CM1, CM2, CM3 models on complex tasks; infrastructure overhead for self-hosting; not yet validated at full SLR-pipeline scale within the current project.

SLR Suitability: Medium. The most viable open-source candidate; recommended for further investigation including domain-specific fine-tuning experiments in subsequent work package tasks.

---

5. Comparative Analysis

5.1 Performance vs. Cost Trade-off

The evaluation reveals a clear trade-off between performance quality and operational cost. The CM1, CM2, CM3 models deliver the highest extraction and summarization fidelity, making them

the most reliable options for SLR tasks requiring nuanced interpretation of scientific methodology. However, these advantages come with API cost dependencies and closed-source constraints.

The OM4 presents the strongest open-source alternative, offering medium performance at near-zero per-token cost, making it the most cost-effective path to a locally deployable, data-sovereign SLR pipeline with fine-tuning potential.

5.2 SLR Pipeline Architecture Recommendation

Based on current findings, the team is evaluating a **tiered deployment architecture** as follows:

1. **Screening tier** (title/abstract): CM3 or OM4 — high speed, low cost, binary classification.
2. **Full-text extraction tier**: CM1 — high accuracy for structured data extraction against predefined schemas.
3. **Summarization and synthesis tier**: CM1 or CM2 — complex multi-document synthesis and report generation.

A parallel fine-tuning track for OM4 on SLR-annotated corpora is planned for WP3, with the intent of reducing dependence on proprietary API services over the project lifecycle.

5.3 Open-Source Accessibility Summary

Model	Licence	Local Deployment	Fine-Tuning Feasibility
CM1	Proprietary	✗	✗
CM2	Proprietary	✗	✗
CM3	Proprietary	✗	✗
OM5	Apache 2.0	✓	✓
OM4	Apache 2.0	✓	✓

---

6. Risks and Mitigation

Risk	Affected Models	Mitigation

API dependency and service availability	CM1, CM2, CM3	Maintain parallel open-source deployment track; architect pipeline for model swappability.
Cost overrun at scale	CM2	Reserve for high-complexity tasks only; use Haiku or Scout for volume screening.
Performance gaps in open-source models	OM4, OM5	Planned domain fine-tuning (WP3); phased integration with human-in-the-loop validation.
Data sovereignty and third-party data processing	CM1, CM2, CM3	Evaluate data processing agreements; assess feasibility of on-premises LLM deployment for sensitive corpora.
Model deprecation or API pricing changes	All API models	Version-lock API integrations; maintain abstraction layer for model substitution.
