

AI for Incidental Findings:

A Comprehensive Guide to Choosing Between NLP, LLMs, and Computational Linguistics

Executive Summary

Managing incidental findings in radiology requires AI systems that can interpret clinical context, not merely detect the presence of certain words. This guide examines three categories of AI commonly used to extract or interpret information from radiology reports—Natural Language Processing (NLP), Large Language Models (LLMs), and Computational Linguistics (CL)—and explains how their underlying architectures shape both **identification accuracy** and the **operational work required to use AI safely in clinical practice**.

Central to this discussion is a concept referred to here as the Validation Burden—a form of operational overhead that is rarely measured explicitly, but widely experienced across health systems deploying clinical AI. It reflects the often-overlooked work required to manually review, verify, and contextualize AI-generated findings before they can support clinical workflows or downstream decisions. Validation burden has become one of the largest hidden costs in clinical AI. Health systems often expect automation but underestimate the hours and staffing required to validate AI outputs before anyone can act on them.

For incidental findings and screening programs—where volumes are high and clinical characteristics are complex—this burden can quickly eclipse expected efficiency gains, increase staffing requirements, and slow patient follow-up, all at significant financial cost to the health system.

Understanding how different AI architectures influence both identification precision and validation burden is essential for AI Governance committees assessing safety, scalability, and total cost of ownership.

Introduction: The Challenge of Incidental Findings

This scenario illustrates the inherent challenge of incidental findings. These findings are unrelated to the reason imaging was ordered, yet they may require surveillance, diagnostic testing, or specialist referral. They are clinically meaningful but operationally easy to overlook.

Across the United States, millions of incidental findings appear in radiology reports each year. Depending on the condition—such as pulmonary nodules, pancreatic cysts, renal lesions, adrenal nodules, or aortic aneurysms—studies suggest that 30–40 percent do not receive recommended evidence-based follow-up. The consequences can include delayed diagnoses, missed early-stage cancers, unnecessary disease progression, and avoidable mortality.

Several characteristics make incidental findings uniquely difficult to manage at scale.



Why Incidental Findings Are Difficult to Track

Several inherent characteristics make incidental findings uniquely challenging for automation and workflow integration:



Incidental findings are not the focus of the exam



Radiology reports are unstructured narrative documents



Clinical significance depends on context not keywords alone



Incidental findings require longitudinal tracking over time



Small details carry large clinical and operational weight

AI offers a potential path forward—but only when it can extract meaning with high precision, interpret context correctly, and minimize the amount of human work required to make outputs safe to act on.

Three AI Approaches Used for Radiology Reports

AI systems used to interpret radiology reports are often discussed as if they are interchangeable. In practice, they are built on fundamentally different architectures that behave very differently when applied to incidental findings and screening programs. These differences directly affect **how accurately findings are identified and how much validation work is required before follow-up can occur.**



Basic Natural Language Processing (NLP)

Basic Natural Language Processing systems rely on techniques such as named entity recognition and dictionary-based matching to identify predefined terms in text. These systems can locate words like "nodule," "mass," "6 mm," or "right upper lobe" and tag them as relevant entities.

Their primary strength lies in speed and scalability. For retrospective analysis, simple filtering, or use cases where approximate identification is sufficient, basic NLP can be effective and computationally efficient.

However, basic NLP does not inherently understand meaning or relationships. It may extract a measurement, an anatomical location, and a descriptor without knowing whether they refer to the same finding. Negation handling is limited, leading to errors such as flagging "no evidence of pulmonary nodule" as a positive finding. Temporal context is also difficult to manage; prior findings may be conflated with current ones, and growth or stability over time is rarely interpreted correctly. Section awareness—distinguishing impressions from history or prior exams—is often incomplete or absent.

As a result, identification accuracy is limited not because terms are missed entirely, but because findings are frequently misattributed, taken out of context, or incorrectly surfaced as clinically relevant.

In clinical workflows, clinicians or navigators must re-read reports to reconstruct meaning and verify whether follow-up is truly required. As volumes increase, this manual reconstruction becomes a major source of validation burden, constraining scalability and slowing patient follow-up.



Large Language Models (LLMs)

Large Language Models generate text probabilistically based on patterns learned from vast corpora. When applied to radiology reports, they can summarize findings, interpret varied phrasing, and respond to complex prompts in ways that appear highly sophisticated.

Compared to basic NLP, LLMs handle linguistic nuance more effectively and can accommodate variability in reporting style. This makes them appealing for narrative summarization, education, or other low-risk informational use cases.

In structured clinical extraction workflows, however, LLMs introduce a different set of challenges. Because they generate language rather than extract it deterministically, they may hallucinate—introducing details not present in the source text—or subtly alter measurements, timeframes, or descriptors. Their reasoning is opaque, making it difficult to trace outputs back to exact source language. Outputs are also non-deterministic, meaning the same input can yield different results across runs.

This variability makes **identification accuracy inherently unstable**, particularly for incidental findings where small differences in size, timing, or wording can change whether follow-up is required at all. As a result, every output must be carefully reviewed to ensure that nothing has been fabricated, omitted, or reinterpreted in a clinically meaningful way. For high-volume incidental findings and screening programs, this level of verification significantly increases validation burden and limits safe operational scaling.



Computational Linguistics (CL)

Computational Linguistics takes a different approach. Rather than predicting language, CL systems analyze syntax, semantics, and relationships directly from source text using explicit linguistic rules, domain ontologies, and structured logic.






CL models designed for incidental findings recognize report sections, identify entities, and explicitly map relationships between measurements, anatomical locations, descriptors, and recommendations. They distinguish current findings from prior ones, interpret temporal language, and assess changes such as growth or stability over time. Every extracted element remains traceable to the source text.

Because CL systems are deterministic, the same input produces the same output every time. There is no risk of hallucination, and outputs can be fully audited—an essential requirement for clinical governance and regulatory defensibility. This architectural precision enables **consistently high identification accuracy**, reducing both false negatives that delay care and false positives that inflate manual review.

This approach does require upfront domain modeling and condition-specific configuration, which can make Computational Linguistics less flexible for broad, open-ended language tasks.

In practice, higher identification accuracy directly reduces validation burden. When extracted information is precise, complete, and correctly contextualized, human oversight shifts from reconstructing meaning to confirming accuracy. For incidental findings and screening programs operating at scale, validation becomes an occasional safeguard rather than a continuous operational requirement.

Validation Burden decreases when:

-  **Extracted information is precise and complete**
Findings are identified with sufficient detail to determine follow-up without re-reading the report.
-  **Clinical relationships are captured accurately**
Measurements, locations, and descriptors are correctly linked to the same finding.
-  **Temporal context is interpreted correctly**
Current findings are distinguished from prior ones, and change over time is understood.
-  **Outputs are deterministic and reproducible**
The same input produces the same result, supporting auditability and trust.
-  **Every data element is traceable to source text**
Clinicians can see exactly where information came from without manual reconstruction

The Validation Burden as an Operational Reality

In clinical AI, accuracy alone is not sufficient. Every AI-generated output that informs patient care must be reviewed, trusted, and acted upon by a qualified professional. Validation burden increases when AI systems miss context, generate ambiguous outputs, conflate past and present findings, or require clinicians to infer intent.

Across AI approaches, the most meaningful operational difference is not whether findings are surfaced, but **how much human effort is required to validate, trust, and act on the output**. In high-volume programs, validation burden—rather than detection capability alone—often becomes the limiting factor for scalability.

Choosing the Right AI Model for Incidental Findings and Screening Programs

Different AI models are appropriate for different tasks:

Basic Natural Language Processing

May be sufficient when the goal is simple keyword search or retrospective analysis and human reviewers can easily add missing context.

Large Language Models

May be appropriate for narrative summarization or educational use cases where outputs do not directly drive clinical workflows.

Computational Linguistics

Most appropriate for use cases where accuracy, traceability, determinism, and longitudinal workflow integration are essential—and where false positives and false negatives materially affect workload and patient outcomes.

Why it Matters

Incidental findings sit at the intersection of clinical nuance and operational scale. While AI can help surface these findings, the architecture behind the model determines how accurately they are identified—and how much human work is required before follow-up can safely occur.

Validation burden makes that tradeoff visible. Understanding where work, cost, and risk ultimately reside after AI is deployed is essential for organizations managing incidental findings and screening programs at scale. Understanding these differences is central to evaluating not just model performance, but long-term safety, staffing impact, and total cost of ownership.

About Eon

Eon is a healthcare technology company focused on supporting health systems in the identification and ongoing management of patients at risk of cancer and other life-threatening conditions. Powered by condition-specific clinical AI, Eon's longitudinal care management platform extracts incidental findings documented in radiology reports and helps ensure patients receive timely, guideline-based follow-up and remain in appropriate surveillance over time.

More than 70 health systems across over 1,200 facilities rely on Eon and its care management services to scale early detection programs, enable earlier diagnosis and treatment, and support sustained patient engagement—outcomes that also carry meaningful financial implications for health systems.