Every Contact Leaves a Trace: Principles for Trustworthy Artificial Intelligence in Medicolegal Death Investigation

Mihir Joshi[1,4*] · **Vikas Kakar**[1,4*] · **Maneesha Pandey, MD**[1,2,3]

[1]Locarda · [2]Forensic Medical, Kansas City, KS · [3]The Forensic Pathologists, Toledo, Ohio
[4]Contributors: NAME Artificial Intelligence Subcommittee of the Strategic Planning Committee
*presenting authors · founders@locarda.com

MDI-Bench
Full methods and results

BACKGROUND

Medical examiner and coroner offices sit where medicine, public health and the justice system meet, and what they write carries consequences for families and in court [1]. Forensic pathology has to define what trustworthy artificial intelligence looks like in its own practice.

The problem

AI already writes clinical documentation, and its failure modes are documented: omitted findings, fabricated content and misattribution that survive review [2, 3]. A system that scores well on general medical benchmarks can still fail at the work itself [4].

Why existing evaluations do not answer it

Documentation benchmarks score against a single reference note [5, 6, 7]; legal work-product [8] and forensic evaluations [9, 10, 11] are examination focused. All assume one correct answer, which fails where informed judgment varies [12, 13, 14, 15, 16].

What we propose: judge an AI-written section against the text that office actually produced, not a generic ideal. Cause of death stays the pathologist's call; what is measured is whether the system carries the dictated findings into that office's style.

Contributions

- A task and a corpus from real casework: 1,486 dictated sections from 230 cases, hand labeled by the five pathologists who dictated them.
- Three scoring axes — lexical fidelity, content sequencing, layout mechanics — by fixed rules, reported separately, never fused into one number.
- Three frontier systems measured against a do-nothing floor, with the office, not the model, as the unit of personalization.

Locarda is an interdisciplinary team spanning engineering, AI research and forensic pathology, with more than 1,000 examinations supported by AI [17, 18, 19].

ONE DICTATION, TWO OFFICES

"Two inch abrasion, left forearm. One inch contusion, right knee. Half inch laceration above the left eyebrow."
The dictation

OFFICE 1
Extremities:
A 2 inch abrasion is present on the left forearm.
A 1 inch contusion is present on the right knee.
Head:
A 1/2 inch laceration is present above the left eyebrow.
Full sentences, grouped by region.

OFFICE 2
- Abrasion, left forearm, 2.0 inches.
- Contusion, right knee, 1.0 inch.
- Laceration, left eyebrow, 0.5 inch.

Telegraphic, one flat list.

FIGURE 1. WHAT MDI-BENCH MEASURES AND HOW IT WAS BUILT

MDI-Bench
Medicolegal Death Investigation Benchmark

4 TEMPLATES, 3 OFFICES

- **Office 1** — 2 templates
- **Office 2** — 1 template
- **Office 3** — 1 template

Every task is defined against its own office's template, not a generic one.

1,486
sections graded

from 230 decedent cases
42 section types
5 pathologists

Can an AI system write an autopsy section the way a particular office writes it: the dictated findings carried across intact, in that office's own style?

HOW THE BENCHMARK WAS BUILT

1 · DICTATE
The pathologist dictates a section; a draft appears.

2 · CORRECT
They edit it until it matches their own style.

3 · ONE GRADED ROW
dictation + section + template → reference

Hand labeled by the pathologists, section by section.
Every reference is text that office actually produced, in its own house style.
Not an annotator's reconstruction.

WHAT IS SCORED · THREE AXES, REPORTED SEPARATELY

Lexical fidelity
Do the words match?
reference: A 2 inch abrasion is present
output: A 2 inch abrasion was noted
Substitutions, drops and additions over aligned words.
scored 0 to 1, no weights

Content sequencing
Is it in the right order?
reference: Extremities → Head
output: Head → Extremities
Reciprocal-best block matching, Spearman footrule.
scored 0 to 1, no weights

Layout mechanics
Is the surface the same?
reference: 1 EVIDENCE 2 ¶ 3 ····text
output: 1 EVIDENCE 2 ··text
Blank lines, indents, tabs, markers, punctuation, units.
scored 0 to 1, no weights

De-identified autopsy reports. The answer for every task is the section that office produced.

FIGURE 2. HOW ONE SECTION IS SCORED

1 · ONE TASK
the de-identified dictation
the section it belongs in
that office's blank template
one completed example from the same office

2 · TWO TEXTS TO COMPARE
WHAT THE SYSTEM WROTE
the candidate section
WHAT THE OFFICE PRODUCED
the reference section

3 · THREE AXES, SCORED SEPARATELY
Lexical fidelity — substitutions, drops and additions over aligned words
Content sequencing — is the material in the same order as the reference
Layout mechanics — blank lines, indents, tabs, markers, punctuation, units

4 · WHAT COMES OUT
Each axis scores 0 to 1. The overall score is their unweighted mean. No axis is weighted above another and none is combined away.
Computed by fixed rules, not by another language model.

Every section is scored a second time with the dictation submitted unchanged.
That baseline shows how much of a score was already in the input before any model touched the text.

Every score on this poster comes out of this pipeline. The axes are computed by fixed rules, not by another language model, and none of them is weighted above another.

METHODS

The task

One dictated section at a time. The system receives the dictation, the section name, that office's blank template and one completed report from the same office, and returns the formatted section. Nothing else: no extra instructions, no worked examples or retrieval. The two documents are the entire specification of the house style.

The corpus

1,486 dictated sections from 230 decedent cases, across 4 report templates from 3 offices, dictated by 5 pathologists. The reference is the section as it entered the record. Median dictation is 209 characters against a median reference of 296; 66% of references are a single block. All text is de-identified.

Scoring

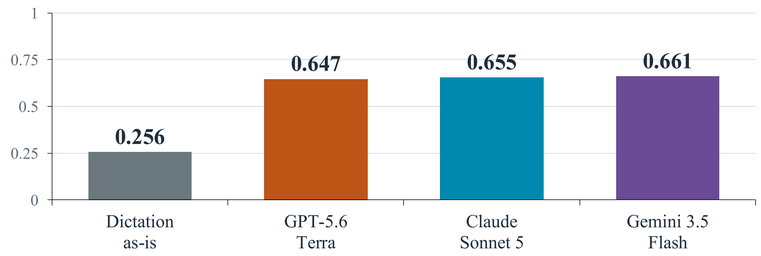
Three axes, each bounded 0 to 1, reported separately and never weighted: lexical fidelity (word alignment: substitutions, omissions, additions), content sequencing (blocks matched reciprocally, scored by Spearman footrule) and layout mechanics (blank lines, indents, markers, punctuation, units). Fixed rules throughout, no language model as judge. The overall score is their unweighted mean.

Three systems, one harness, one prompt shape: 4,458 graded rows, plus a do-nothing floor of the dictation left unchanged.

RESULTS

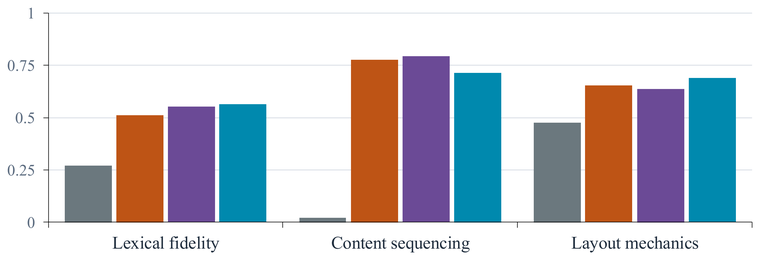
Dictation · GPT · Gemini 3.5 · Claude

Figure 3. Overall score, mean of the three axes (0 to 1)



The three models sit within 0.014 of each other and all three clear the dictation submitted unchanged by a wide margin. This is not a vendor ranking.

Figure 4. Each axis scored separately, against the dictation unchanged (mean, 0 to 1)



Ordering is where the models add most: the raw dictation scores 0.02 and they reach 0.71 to 0.79. Wording is the unsolved axis, 0.51 to 0.56.

CONCLUSIONS

Three things to check when AI writes an autopsy section: the office's words, the order of the material, and the layout.

All three frontier models land near the same score and far above the raw dictation. They plateau together, so the limit is not model quality.

The gap is specific: getting the order right is largely handled, writing in the office's own words is not.

Closing it means adapting to the office, not scaling the model. Until then a draft still needs editing for wording even when its structure is right.

LIMITATIONS & FUTURE DIRECTIONS

Limitations

- **One office supplies 76%** of the sections.
- **Sequencing is the thinnest axis,** applying to about a third of sections, so it is the least stable.
- **The axes measure agreement with the reference text,** not whether a pathologist would accept the draft.
- **One condition, one time point:** a generic harness, three models, a single de-identified corpus.

Future directions

- **Section routing, instruction following** and mid-report edits that preserve untouched content.
- **Medical record review, handwriting and scanned documents,** and physician rated reviewability across offices.

REFERENCES & ACKNOWLEDGMENTS

With thanks to the participating medical examiner offices.

1. National Association of Medical Examiners. (2025). Forensic autopsy performance standards (5th ed.). Louisville, KY: NAME.
2. Topaz, M., et al. (2025). Beyond human ears: risks of artificial intelligence scribes in clinical practice. npj Digital Medicine, 8.
3. Atiku, S., et al. (2026). Socio-technical mechanisms shaping AI scribe documentation failures. Journal of Evaluation in Clinical Practice.
4. Gong, E. J., et al. (2025). The knowledge-practice gap in clinical large language models. Journal of Medical Internet Research, 27, e84120.
5. Arora, R. K., et al. (2025). HealthBench: evaluating large language models towards improved human health. arXiv:2505.08775.
6. Yim, W., et al. (2023). ACI-Bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Scientific Data, 10.
7. Ben Abacha, A., et al. (2023). Overview of the MEDIQA-Chat 2023 shared tasks on the summarization of doctor-patient conversations. 5th Clinical NLP Workshop.
8. Harvey. (2024). BigLaw Bench: measuring artificial intelligence on real legal work product. Technical report.
9. Aydogan, H. C., et al. (2025). Evaluation of ChatGPT-4 generated forensic autopsy reports. Egyptian Journal of Forensic Sciences, 15, 30.
10. Fu, E., et al. (2026). Large language models as consultants for cause of death determination. Frontiers in Artificial Intelligence, 9, 1752043.
11. Sohail, A., et al. (2026). Benchmarking multimodal large language models for the forensic sciences. IEEE Access, 14.
12. Davani, A. M., et al. (2022). Dealing with disagreements: looking beyond the majority vote in subjective annotations. TACL, 10, 92-110.
13. Gordon, M. L., et al. (2022). Jury learning: integrating dissenting voices into machine learning models. CHI '22.
14. Castricato, L., et al. (2024). PERSONA: a reproducible testbed for pluralistic alignment. arXiv:2407.17387.
15. Palm, E., et al. (2025). Assessing the quality of AI-generated clinical notes. arXiv:2505.17047.
16. Zhou, K., et al. (2025). From feedback to checklists: grounded evaluation of AI-generated clinical notes. EMNLP 2025 Industry Track.
17. Pandey, M., & Joshi, M. (2024). Leveraging artificial intelligence and augmented reality to enhance autopsy workflows in forensic pathology. NAME Annual Meeting.
18. Joshi, M., Pandey, M., Kakar, V., & Pradhan, A. (2025). Reimagining autopsy documentation using artificial intelligence. NAME Annual Meeting, Louisville, KY.
19. Joshi, M., Kakar, V., Pandey, M., & Pradhan, A. (2026). Building the artificial intelligence infrastructure for forensics. AAFS Annual Scientific Conference, New Orleans, LA.