Nabla

Dec 2024

# Raising the bar for AI-powered clinical documentation

A deep dive into Nabla's evaluation methods, showcasing best practices in AI-driven documentation with a focus on accuracy, adaptability, and transparency.


**Written by**

Samuel Humeau - Lead Machine Learning Engineer

**Contributions from**

Clément Baudelaire - ML Senior Product Manager

Ronan Sangouard - Machine Learning Engineer

# Introduction

Electronic health records (EHRs) have revolutionized healthcare by simplifying access to patient information. However, they come with a significant trade-off: an overwhelming documentation workload for clinicians. Primary care physicians, for example, often dedicate up to two hours daily to EHR tasks, much of it during personal time–commonly referred to as "pajama time." Nabla addresses this challenge by generating clinical documentation directly from conversations between patients and practitioners, seamlessly integrating with the EHR.

Evaluating AI-generated clinical notes presents unique challenges. Given the variability of notes, robust assessment requires a combination of quantitative metrics and clinician-led reviews to uphold the standards of accuracy, clarity, and reliability critical for patient safety. Nabla prioritizes transparent and rigorous evaluation processes to ensure that its AI-generated notes meet these high standards before deployment. This whitepaper outlines Nabla's approach, showcasing its robust evaluation methods and commitment to advancing best practices in AI-driven documentation.

# Executive Summary

## Committed to alleviating clinician burnout

Nabla is dedicated to reducing the burden of clinical documentation, supporting clinician well-being and enhancing patient care.

## Maintaining transparency and quality

Through detailed monitoring and real-time metrics, Nabla ensures its systems deliver consistent quality while providing organizations with tools for clear oversight.

## Prioritizing feedback

Nabla actively incorporates clinicians feedback to continually refine and improve its systems.

## Delivering specialized capabilities

With advanced speech-to-text, customizable note generation, and ICD-10 coding, Nabla remains committed to supporting clinicians with reliable and adaptable solutions.

## Ensuring rigorous evaluation and safe rollout

Nabla commits to thorough testing of every system update, combining automated evaluations with expert reviews to guarantee accuracy and reliability in AI-generated notes. Updates are introduced progressively and safely.

# 1.

# Quality and monitoring High level process

From early stage development to post-deployment monitoring, iterations on Nabla's system follow a rigorous quality-centric process.

# 1.

## Quality and monitoring - High level process

## Overview of Nabla's documentation system

Nabla's system, from a higher standpoint can be decomposed in 3 successive parts:

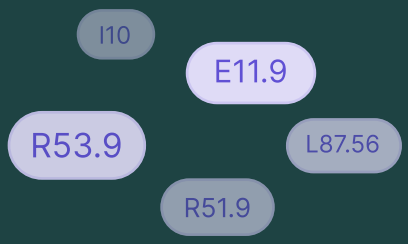
### Speech to text

The speech-to-text brick converts the audio of the medical encounter into a textual version.

### Note generation engine

From the textual version of the encounter, and the patient context from the provider's EHR, the note generation engine produces a medical note.

The note generation engine is highly customizable and offers settings accounting for a wide range of provider preferences.




### ICD-10 coding engine

From the note generated, Nabla deduces which problems should be added to the EHR's problem list and visit diagnosis.

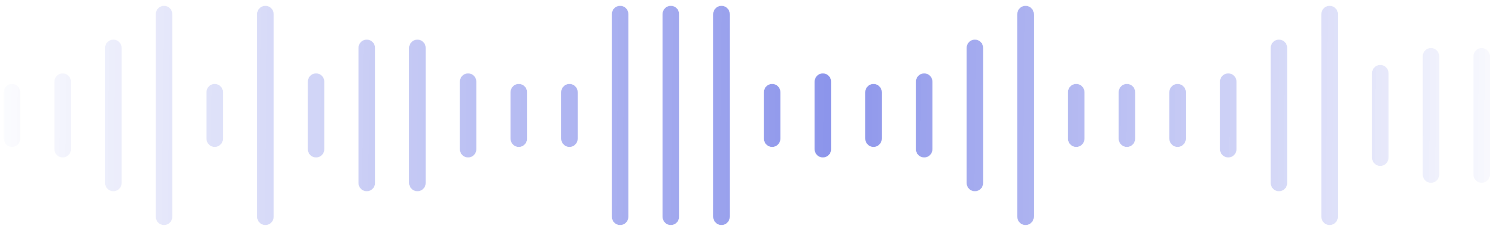
Each of these systems are comprised of multiple sub systems and are evaluated both individually and end-to-end.

# 1.

## Quality and monitoring - High level process

## Automatic metric based development

Every change to Nabla's system or components undergoes rigorous automatic evaluation using a robust dataset of medical encounter audio, gold-standard transcripts, comprehensive note quality tests and expert-reviewed ICD-10 codes. Automated metrics are used as a reliable proxy for final documentation quality, enabling rapid iteration without compromising our high standards.



# 1.

## Quality and monitoring - High level process

### Side by side expert evaluation

Before each major deployment in production, we ask a team of expert medical scribes to evaluate whether the candidate system we want to deploy fares better than the existing one on a large variety of criteria. In particular in the case of the note generation component, the setup is as follow.

SITUATION

Audio, transcript, patient context from EHR, settings required by provider.

NOTE 1

Note generated by current production.

NOTE 2

Note generated by the candidate system.

For each section of the note, the expert is asked to answer precise questions detailing why Note 1 or Note 2 is preferred. Evaluations are aggregated into a sort of “score sheet” which determines whether we should deploy the candidate system or not.

### Example of score sheet showing the details of evaluation criteria on “History of present illness” section.

| EVALUATION CRITERIA | RESULT | WINS | TIES | LOSSES |
|---|---|---|---|---|
| All symptoms, problems and concerns are reported | SIGNIFICANT WIN | 41% | 35% | 23% |
| Absence of symptoms are reported | SIGNIFICANT LOSS | 6% | 81% | 13% |
| Previous history about these symptoms is reported | SIGNIFICANT WIN | 21% | 74% | 4% |

# 1.

## Quality and monitoring - High level process

### Progressive rollout

If the candidate model is validated during the side-by-side evaluation, Nabla employs a progressive rollout strategy for each major update. It means that a certain percentage of users will experience the new model while the majority of our users will be served by the current model. This phased approach enables us to closely monitor the update's impact in a real-world setting.

During this initial rollout, we gather data on performance metrics, user feedback, and potential issues. The number of edits made in the Nabla interface is an excellent passive indicator of quality. We do compute the number of characters added or removed to have a better sense of the depth of these edits. We also closely monitor star-rating and detailed feedback that clinicians using Nabla report. Once we confirm stability and consistent quality with no significant issues, we proceed with a full rollout to all users.

### Monitoring

**Tracking trends at scale**

To ensure continuous improvement of our documentation system, Nabla monitors the scale and type of edits made within each section of generated notes. By analyzing these edits, Nabla can identify trends that reveal specific areas needing further refinement. With a customization-first approach, Nabla aligns to each provider's preferences through a comprehensive range of settings. Edit tracking enables us to pinpoint combinations of settings that may benefit from additional development.

**Empowering our users with custom feedback**

Nabla empowers clinicians to provide direct, custom feedback on the documentation our system generate. Through the user interface, clinicians can provide detailed insights on accuracy, relevance, and usability, allowing us to refine our models based on real-world use. This feedback loop not only helps us enhance the quality of note generation but also fosters a collaborative relationship with our users.

# 1.

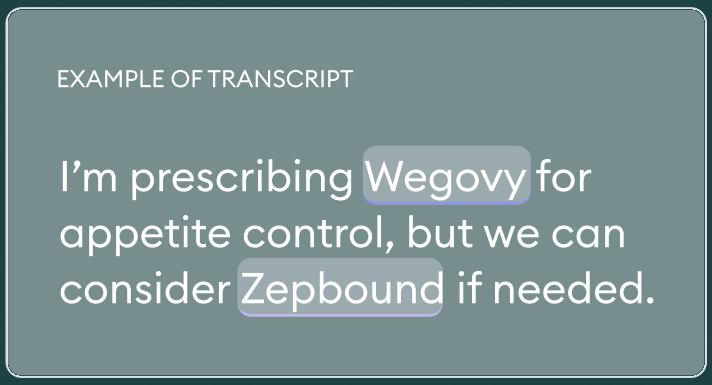
## Quality and monitoring - High level process

Here are some notable advancements that were shaped by valuable user feedback.

### New medication recognition

Wegovy (approved in 2021) and Zepbound (2023) are newer medications.

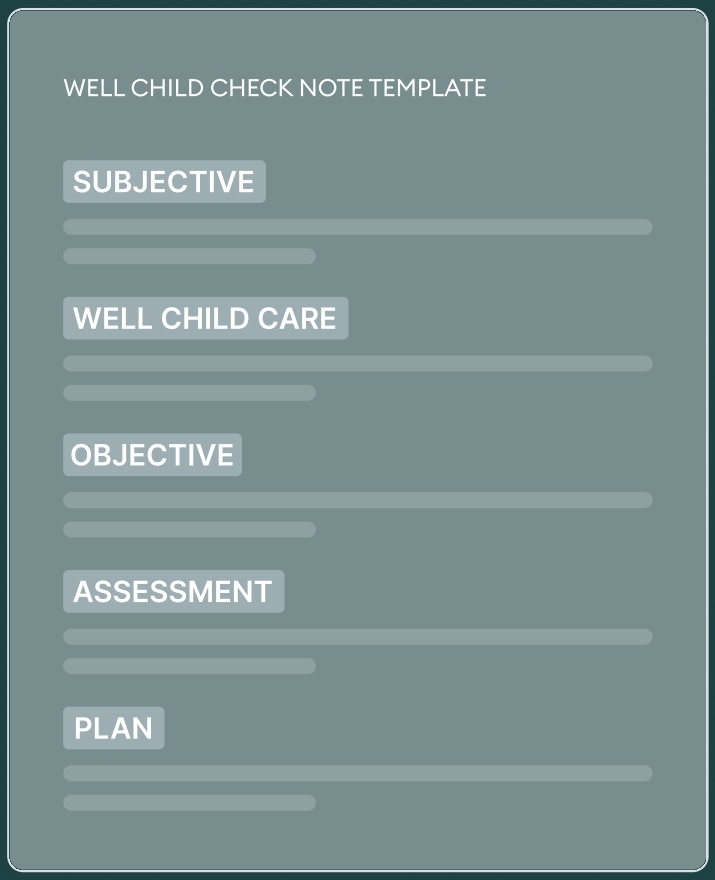
Initially flagged by doctors due to recognition errors, these were quickly resolved by updating our training set.




### Customization per specialty

Pediatric practitioners separate the Well Child Check section from the rest of the subjective section.

This option is now available to all users, alongside other specialty-specific settings.




### Interfacing video calls

A number of practitioners used separate video calls software to address their patient which made audio capture challenging.

Nabla now provides a tool (Nabla Echo) that allows great quality of video calls.

# 2.

# Leading the way on medical speech-to-text

Nabla's clinical documentation generation stems from conversation between a patient and a clinician. Our proprietary speech-to-text transforms this conversation into a written text.

# 2.

## Leading the way on medical speech-to-text

High transcription accuracy is essential, as it directly impacts the final documentation quality.

Nabla relies on a dataset of 7,000 hours of medical conversations, that have been transcribed by professional medical scribes.

## Key metrics

### Word Error Rate

The word error rate loosely measures the number of words wrongly translated.

It is computed as the minimum number of insertions, deletions, and substitutions of words needed to match a transcribed text with the reference text, divided by the total number of words in the reference.

| | |
|---|---|
| **Total number of words in the reference** | 12 words |
| **Changes needed** | 4 changes |
| **Word Error Rate (WER)** | 4/12 = 0.334 |

REFERENCE

I see you are on Metformin these days, for your diabetes, right?

TRANSCRIPT

I saw your Metformin these days, for your diabetes, right?

CORRECTION

I ~~saw~~ see ~~your~~ you are on Metformin these days, for your diabetes, right?

# 2.

## Leading the way on medical speech-to-text

## Key metrics

### Medical Recall

The medical recall focuses on the core terms that will be crucial to establish the documentation.

We define medical term as any of:

- Condition or symptom or sensations
- Medical procedure and tests
- Anatomic terms
- Medication
- Medical devices
- Medical specialty or procedure

It is calculated by computing the percentage of medical terms effectively contained in the transcription.

MEDICAL TERMS

I saw your Metformin these days, for your diabetes, right?

| | |
|---|---|
| **Medical terms** | 2 terms |
| **Medical recall** | 2/2 = 100% |

## Comparative results on one of our internal benchmark

| | WORD ERROR RATE | MEDICAL RECALL |
|---|---|---|
| NABLA | 11.8% | 97.1% |
| WHISPER LARGE V3 | 16.7% | 91.1% |
| GOOGLE SPEECH MEDICAL | 21.7% | 89.4% |

# 3.

# Setting up automatic note generation evaluation

In order to iterate quickly on its note generation engine, Nabla designed a framework to obtain quality metrics during development within minutes.

# 3.

## Setting up automatic note generation evaluation

Nabla relies on a dataset of 50 000 production situations, which have been explicitly shared with Nabla by the practitioners. A production situation contains the current transcript of the encounter, the context of the patient we know from the EHR and settings selected by the clinician.

Typically during iterations Nabla uses a representative subset of 400 production situation. For these situations, careful tests have been crafted and will be applied to the generated notes.

Some of them are programmatic and do look like regular Unit Tests. For example, a test might check, “Is the title longer than 200 characters?”

Most of them rely on a call to a LLM. Content tests for instance, typically are of the form of a simple question regarding the note and an expected answer.
For example: “Does the note contain information XXX?” The expected answer is "yes". This question is prompted to a LLM during the evaluation, and the test is passed if the answer is the one expected.

Evaluating hallucinations (elements that are not in the transcript nor in the patient context but were still generated in the note) or translocation (elements that belong to another section of the note), requires a more complex setup.

A first pass is done to split the generated note into atomic pieces of information.

EXAMPLE

Patient is currently on Metformin for his diabetes.

CAN BE SPLIT ON 3 ATOMIC FACTS

- Patient takes Metformin
- Patient has diabetes
- Metformin is used for diabetes.

For each fact we prompt a LLM to check whether the fact could be deduced from the transcript and patient context (when checking hallucinations) or whether it is located in the right section (when evaluating translocation).

# Conclusion

Nabla's commitment to enhancing clinical documentation reflects a vision of reducing the administrative burden on clinicians while maintaining the highest standards of quality and safety. By focusing on rigorous evaluation, progressive rollouts, and active clinician feedback, Nabla ensures its systems are reliable, adaptable, and aligned with real-world needs.

Through transparent processes, advanced customization, and specialized capabilities, Nabla is not only meeting the demands of modern healthcare but also empowering clinicians to spend more time where it matters most–with their patients. This whitepaper underscores Nabla's dedication to setting a benchmark for trustworthy and efficient AI-driven clinical documentation.

## Contact Us

contact@nabla.com

linkedin.com/company/nablahq

x.com/nabla_ai

# Nabla