When Is a Claim Worth Making?

ToothFairy 4 Challenge

Viktor Buzovsky
viktor.buzovsky@vut.cz

The Faculty of Electrical Engineering and Communication, Brno University of Technology, Czechia

2026

Contents:

1 Diagnosis before design
2 The system
3 System optimization
4 Results
5 Conclusions

Diagnosis before design

the reasons behind our approach

What we tried that did not work for us

- **A 3D medical VLM**
- **Segmentation model with LLM to write the findings**

The problem

On 625 volumes the model learned the style of the reports but not their evidence, so it kept stating findings the image never supported.

What we have built

A segmentation model, deterministic rules that turn the mask into a list of findings, and fixed templates that render them as the report.

Two measurements that reframed the target

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

1. Captioning metrics barely see the image

A **fixed report** (the same 69 words for every patient, no image input) scores METEOR **0.330**.
A second report on the same scan scores **0.357**.

2. Reports are selective, so overlap between them is partial

Scoring one report against another on the same scan gives RadFact F1 **0.525**.
Both reports are correct. Some describe slightly **different findings** of the same case.

The same scan, described by two reports

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Report A

The maxilla is not included in the scan volume.
The condyles are not included in the scan volume.
Absence of tooth 46.
Tooth 48 is impacted in bone and has a horizontal axis.
Absence of tooth 38.
Tooth 37 shows endodontic treatment and a periradicular radiolucent lesion.
The mandibular canal has a regular course.

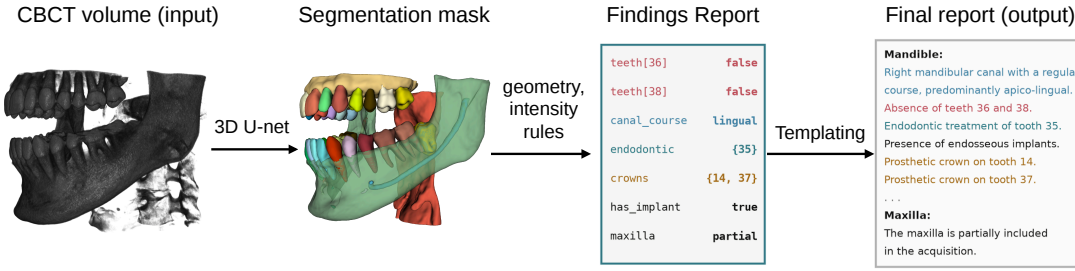
Report B

Mandibular CT including within the acquisition volume the mandibular body and excluding the coronoid and condylar processes bilaterally. Absence of teeth 46 and 38. Endodontic treatment of tooth 37 reaching the apices. **Restoration on the crown of 36. Mandibular canals with a predominantly buccal course bilaterally.** Area of osteorarefaction at the level of tooth 37, **extending to the distal crown of 36 and eroding the lingual wall**, consistent with a likely new formation with a cystic periapical appearance. Tooth 48 semi-impacted with horizontal mesioangular orientation and **not in direct contact with the respective mandibular canal, which courses infero-buccally**.

Scoring B against A: precision $6/13 = 0.46$, recall $6/9 = 0.67$, **F1** $= 0.55$.
Over the 30 cases carrying two reports, the mean is **F1** $= 0.525$. Bold text is what the other report does not state.

The system

The proposed system



- No language model in the whole pipeline. A template is used only when its specific field (finding) is set.
- Every sentence in the final report is produced by a rule you can read.

The segmentation model

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

3D residual U-Net, nnU-Net-style configuration

- 7 levels, shrinks the volume and deepens the features, 32 $\rightarrow$ 320 channels, with residual blocks at every level, then builds back with a decoder.
- patch $120 \times 200 \times 200$ at **0.3 mm** isotropic
- **78 classes**: jaws, canals, sinuses, 32 teeth, 32 pulp chambers, crowns, bridges, implants

Sliding-window inference to fit the container's VRAM and 900s budget.

Dice against ToothFairy3 labels

structure	Dice
mandible	0.987
mandibular canals	0.871
teeth	0.870
pulp chambers	0.855
prosthetic crowns	0.595
all 77 classes	**0.818**

on 30 scans held out from training

Creating the findings record

1. **A label presence:** absent teeth, crowns, bridges, implants
2. **Geometry of the labels:** canal course, canal within 2 mm of a molar, impaction from occlusal reach, how much jaw is in frame
3. **Intensity inside a label:** endodontic treatment when pulp $p_{90} \geq 1.4\times$ bone p_{95}, restoration when tooth $p_{99} \geq 2.2\times$ bone p_{95}

Densities are ratios to **the patient's own jaw bone**, never absolute: CBCT has no shared intensity scale.

18 fields for the **findings record** $\rightarrow$ 22 templates turn it into the **final report**.

77 % of reference statements fall into categories this vocabulary can express. (the missing **23 %** are mostly rarely stated findings)

System optimization

When is a claim worth making?

We can detect tooth filling from intensity inside the tooth label. Such claim is confirmed by the reference about **half** of the time.

Should the system say it?

- Each stated claim the reference does not confirm reduces **precision**.
- Each withheld claim the reference does state reduces **recall**.
- F1 weights the lower of the two more heavily, and for us that is **recall**.

$$\mathrm{F1}(\bar{P}, \bar{R}) = \frac{2\bar{P}\bar{R}}{\bar{P} + \bar{R}}, \qquad \bar{P} = 0.470, \quad \bar{R} = 0.314$$

The break-even entailment rate q^*

Adding a claim entailed with probability q moves both means: it **costs precision** and **buys recall**.

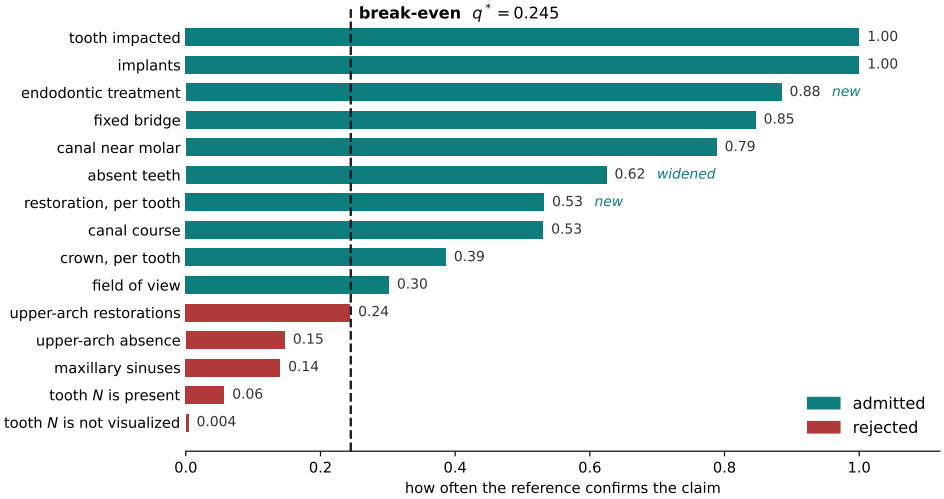
$$\Delta\bar{P} = \frac{q - \bar{P}}{n}, \qquad \Delta\bar{R} \approx \frac{q}{m}, \qquad \frac{\partial \mathrm{F1}}{\partial \bar{P}} = \frac{2\bar{R}^2}{(\bar{P}+\bar{R})^2}, \qquad \frac{\partial \mathrm{F1}}{\partial \bar{R}} = \frac{2\bar{P}^2}{(\bar{P}+\bar{R})^2}$$

The claim is worth making when the resulting change in F1 is positive:

$$\frac{2\bar{R}^2}{(\bar{P}+\bar{R})^2} \cdot \frac{q - \bar{P}}{n} + \frac{2\bar{P}^2}{(\bar{P}+\bar{R})^2} \cdot \frac{q}{m} > 0 \quad \Longrightarrow \quad \boxed{q > q^* = \frac{\bar{P}}{1 + \bar{P}^2 n / \bar{R}^2 m}}$$

- n, m: phrases per case, ours and the reference's.
- baseline $q^* = 0.21$, final $q^* = \mathbf{0.245}$ as the system improved.
- A claim the reference confirms **one time in four** still raises the score.

Filtering candidate findings by q^*



Admit a candidate when $q > q^*$. The line is the threshold at the final operating point, and each number is how often the reference confirms that claim.

Results

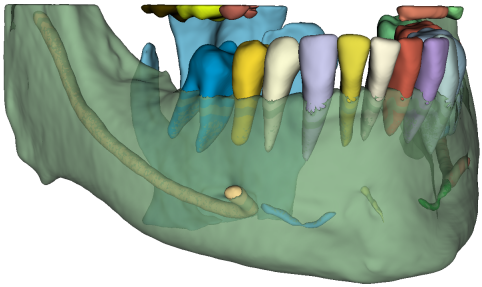
The gain, and the reference point

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

	$\bar{P}$	$\bar{R}$	**F1**	BLEU-4	METEOR
Before calibration	0.470	0.314	0.376	0.208	0.384
After	0.519	0.421	**0.465**	0.226	0.404
Second report	0.514	0.536	*0.525*	0.179	0.357

- Local judge: `qwen3.6-35b-a3b` (official: `GPT-5.6-Luna`)
- The official score is computed on **hidden cases** (different samples).

Output example



Case P035. RadFact on this case:

P 0.75 / R 0.90 / **F1** 0.82

- stated in the reference too
- not stated in the reference

Mandible:

Mandibular CT including within the acquisition volume the mandibular body and excluding the coronoid and condylar processes bilaterally.
Right mandibular canal with a regular course, predominantly apico-lingual.
Left mandibular canal with a regular course, predominantly apico-lingual.
Absence of teeth 47 and 48.
Partial edentulism.
Endodontic treatment of tooth 35.
Endodontic treatment of tooth 37.
Endodontic treatment of tooth 45.
Prosthetic crown on tooth 35.
Prosthetic crown on tooth 37.
No definite osteolytic or osteocondensing lesions.

Maxilla:

Not included in the acquisition volume and not assessable.

Mandibular condyles:

The mandibular condyles are not included in the acquisition volume and are not assessable.

Conclusions

Conclusions

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

1. A **deterministic, segmentation-grounded** pipeline reaches F1 0.465, against about 0.525 for a second clinician's report.
2. Every statement is **auditable**: each sentence traces to a rule in the code, and most of them to a measurement on the mask.
3. A **baseline to beat** and a base to build on: language models shall beat this deterministic approach in order to justify their opacity and hallucination possibility.

Code: `https://github.com/vbuzovsky/TF4_Segmentation-Grounded-Baseline`

Thank you. Questions?

Our team at BUT and UHB:

Zdenek Smekal
smekal@vut.cz

Kamil Riha
rihak@vut.cz

Viktor Buzovsky
viktor.buzovsky@vut.cz

(Department of Oral and Maxillofacial Surgery)

Vojtech Perina
perina.vojtech@fnbrno.cz

Ondrej Liberda
liberda.ondrej@fnbrno.cz

Petr Pokorny
pokorny.petr@fnbrno.cz