Artificial Intelligence & Cybersecurity

Generative AI in Healthcare: Applications and Evaluation of Effectiveness

Candidate:
Lorenzo Zanolin

Supervisor:
Prof. Giuseppe Serra
Co-Supervisor:
Prof. Jan Steinbrener

Dec 13, 2024

Background & Motivations

Background

- Generative AI *goal*: derive a probability distribution from data to generate new synthetic data.
- *Masked Multi-Head Self Attention*: captures global dependencies, process sequences in parallel and generates embeddings that improve performance on a wide range of tasks.

Background & Motivations

Background

- Generative AI *goal*: derive a probability distribution from data to generate new synthetic data.
- *Masked Multi-Head Self Attention*: captures global dependencies, process sequences in parallel and generates embeddings that improve performance on a wide range of tasks.

Motivations

- Generative AI can be integrated within workflows to help clinicians.
- Absence of a complete evaluation framework in the literature $\rightarrow$ new framework that combines human feedback and metric based feedback.

Platform & Components

Salesforce

- Enables companies to manage relationships with customers, prospects and employees through customizable interaction rules.
- Integrates with modules such as Health Cloud and Einstein.

Platform & Components

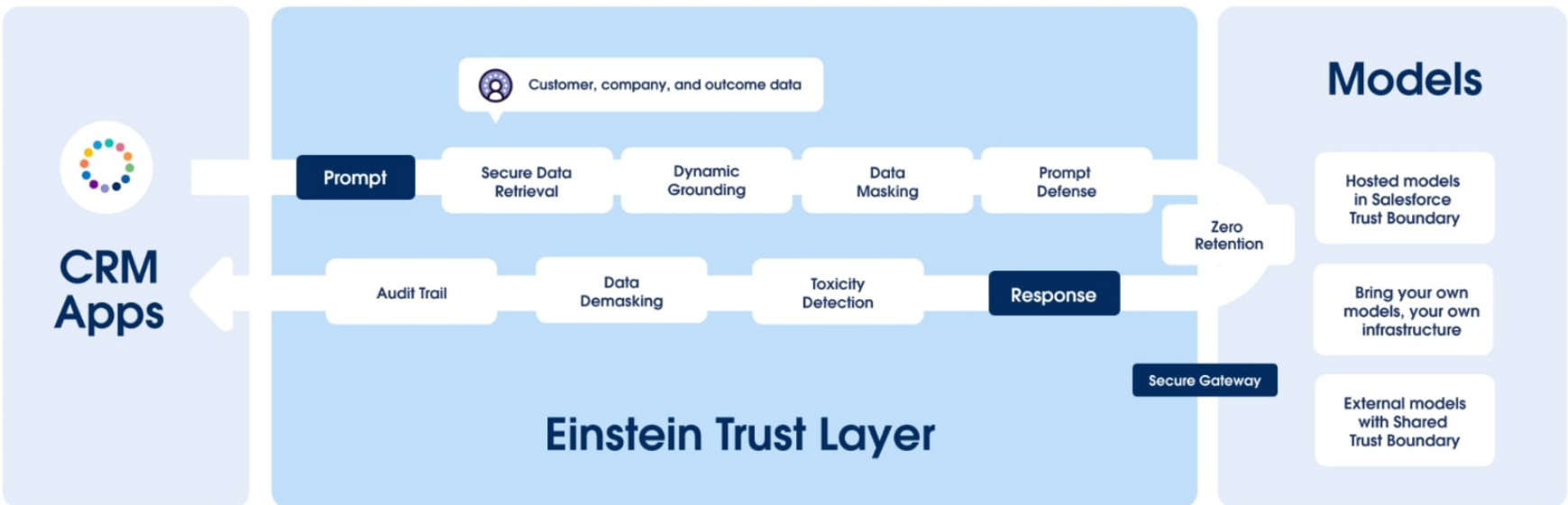
Salesforce

- Enables companies to manage relationships with customers, prospects and employees through customizable interaction rules.
- Integrates with modules such as Health Cloud and Einstein.

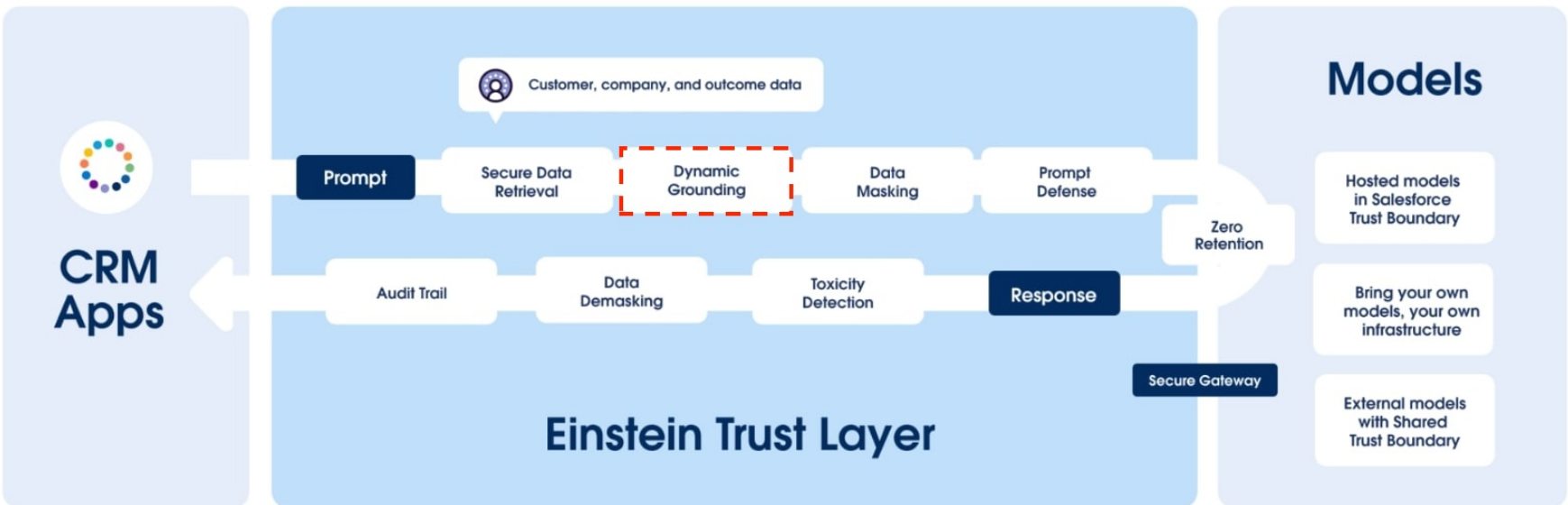
Einstein 1

- Enables the use of genAI models within Salesforce.
- Allows for the creation of reusable prompts with database integration (RAG).
- Trust Layer framework to ensure data security during information exchange with LLMs.

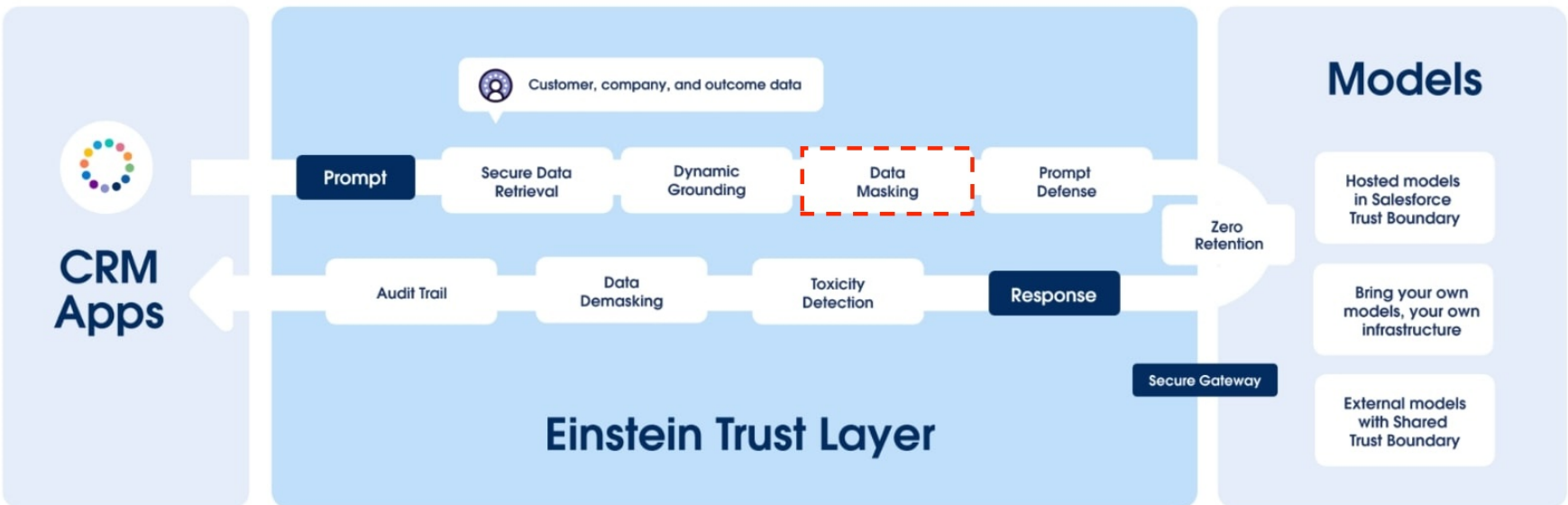
Einstein Trust Layer



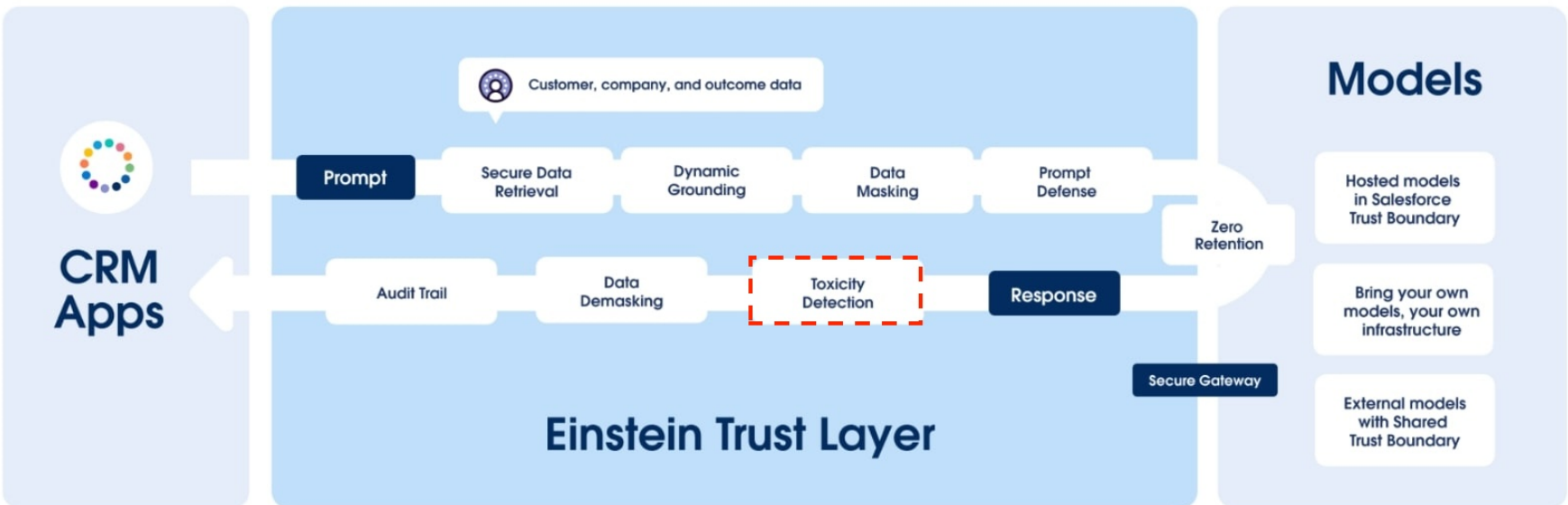
Einstein Trust Layer



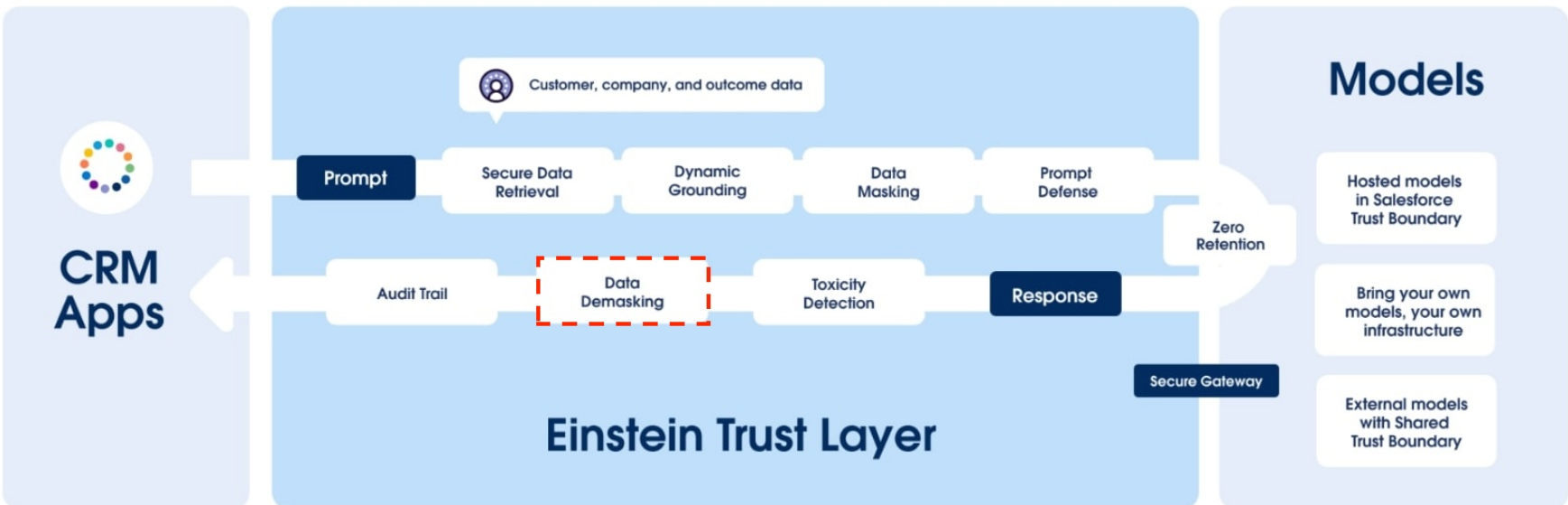
Einstein Trust Layer



Einstein Trust Layer



Einstein Trust Layer



Use Case

Health Clinic

- Application intended for healthcare clinics that manage numerous patients and offer a wide range of services.

- Tasks to be implemented with Copilot:

Use Case

Health Clinic

- Application intended for healthcare clinics that manage numerous patients and offer a wide range of services.

- Tasks to be implemented with Copilot:
  - Patient Summary

Use Case

Health Clinic

- Application intended for healthcare clinics that manage numerous patients and offer a wide range of services.

- Tasks to be implemented with Copilot:
  - Patient Summary
  - List Possible Problems

Use Case

Health Clinic

- Application intended for healthcare clinics that manage numerous patients and offer a wide range of services.

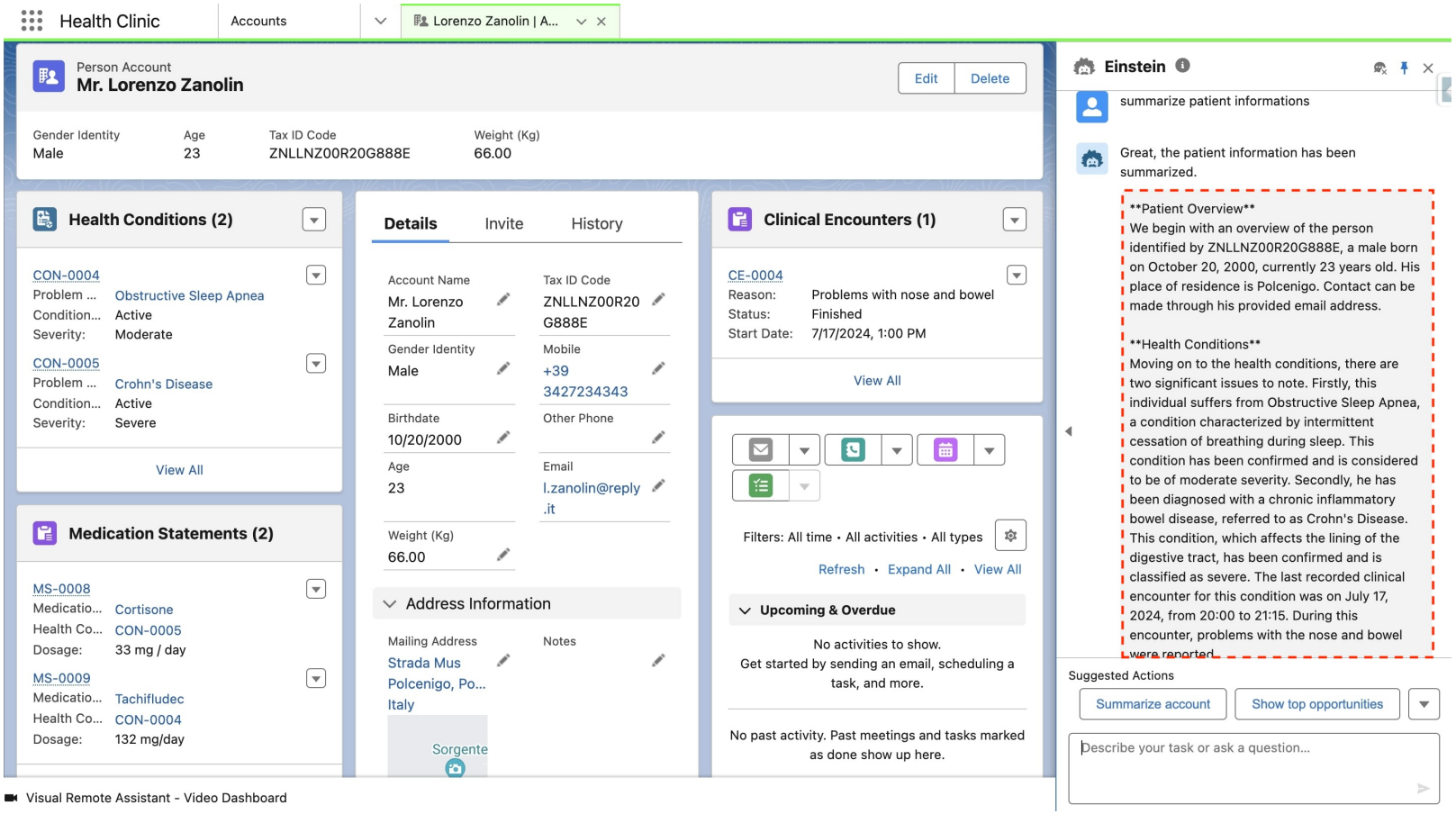
- Tasks to be implemented with Copilot:
  - Patient Summary
  - List Possible Problems
  - Send Visit Details

Patient Summary

Invoked by the doctor before receiving the patient: quick and detailed overview of all the patient's conditions.

Patient Summary

Invoked by the doctor before receiving the patient: quick and detailed overview of all the patient's conditions.

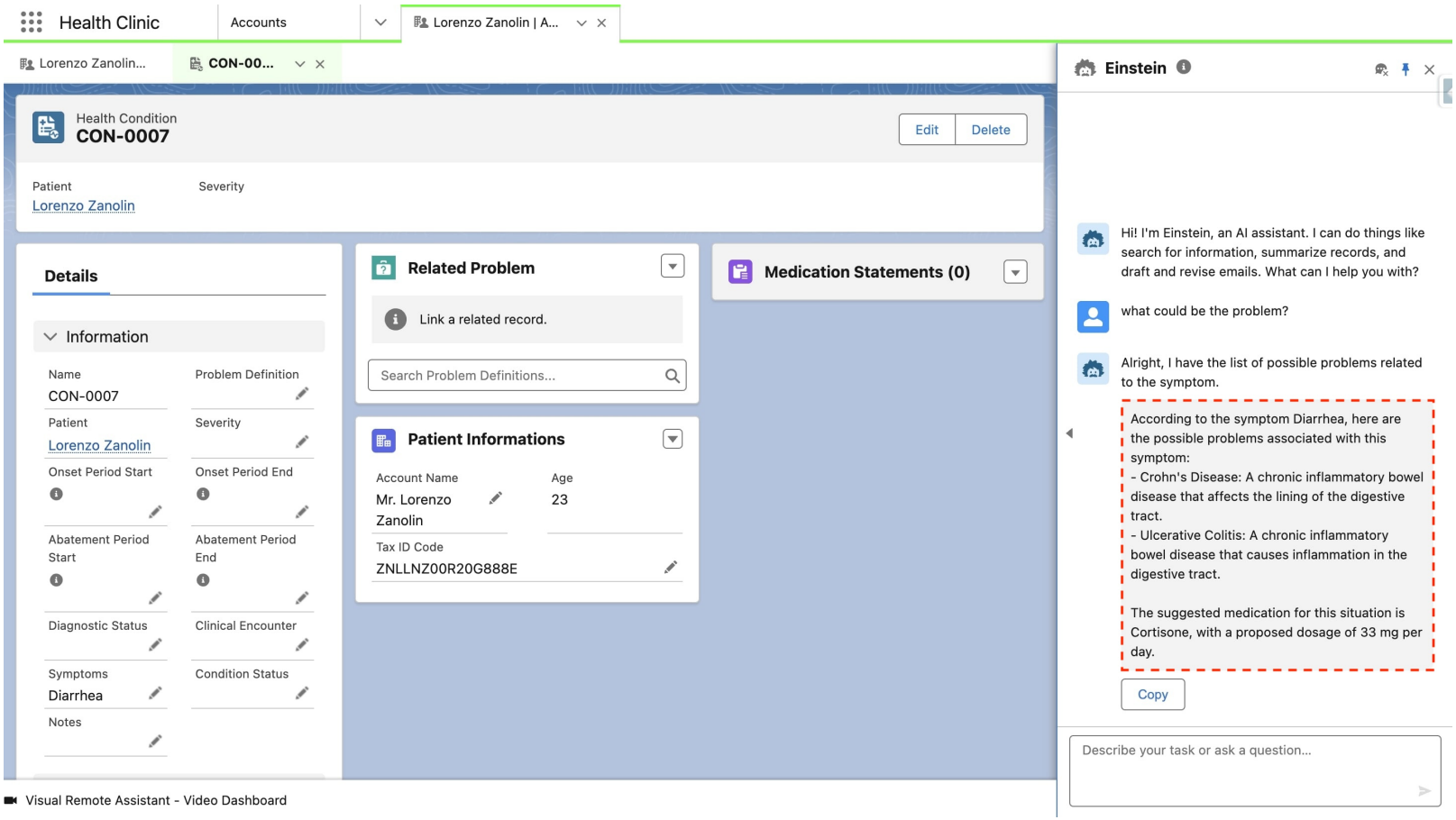


List Possible Problems

Identifies potential conditions based on the patient's symptoms, providing a list of possible diagnoses and medications, with dosage calculations for of them.

List Possible Problems

Identifies potential conditions based on the patient's symptoms, providing a list of possible diagnoses and medications, with dosage calculations for of them.

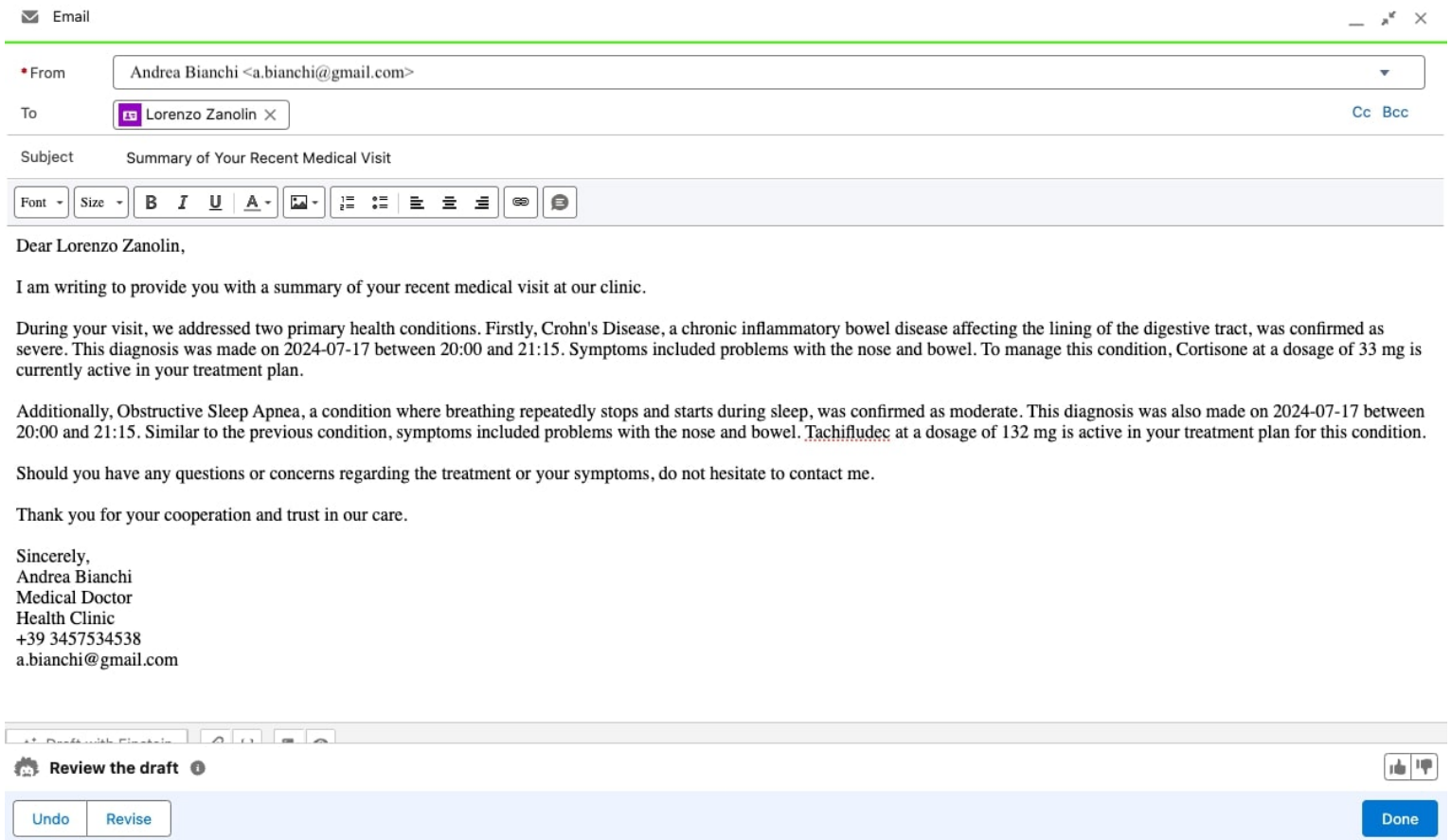


Send Visit Details

Creates a draft email summarizing the patient's latest Clinical Encounter for the doctor to review and edit.

Send Visit Details

Creates a draft email summarizing the patient's latest Clinical Encounter for the doctor to review and edit.



Evaluation Framework

For each task, the output of each model was evaluated in comparison to the clinician's notes.

Evaluation Framework

For each task, the output of each model was evaluated in comparison to the clinician's notes.

Automatic Evaluation:

- Metrics: ROUGE, BLEU, METEOR, Word2Vec and BERTScore.

Evaluation Framework

For each task, the output of each model was evaluated in comparison to the clinician's notes.

Automatic Evaluation:

- Metrics: ROUGE, BLEU, METEOR, Word2Vec and BERTScore.

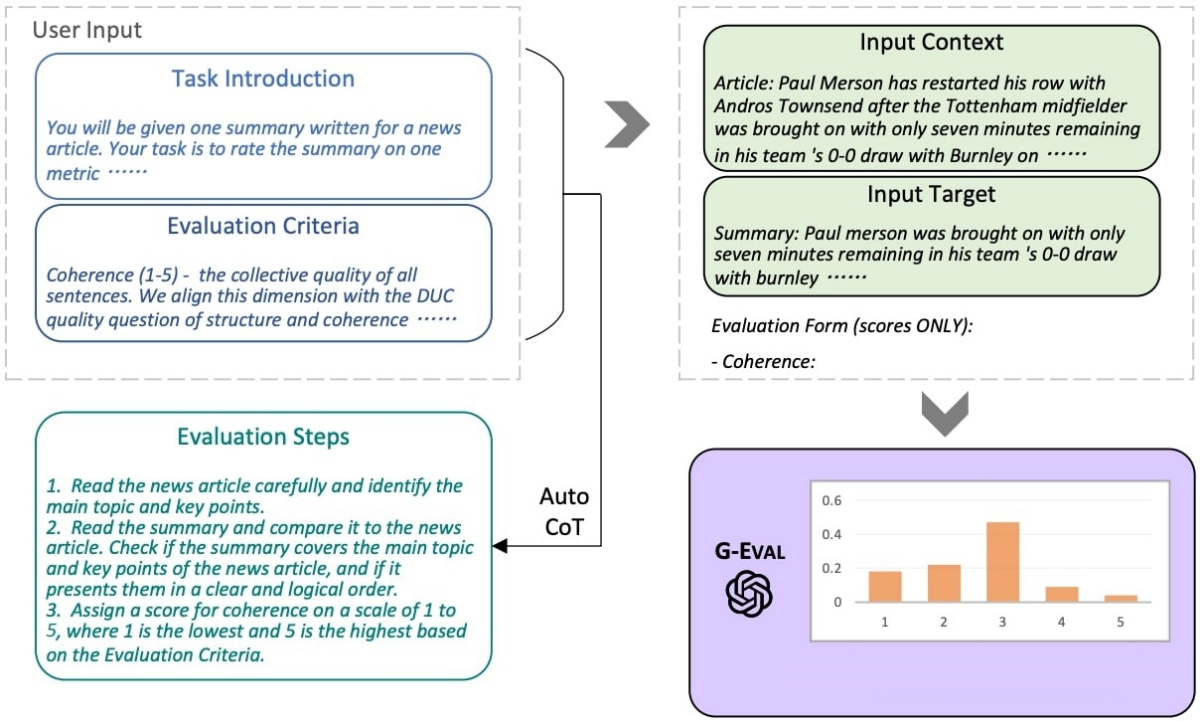
Human Evaluation:

- A sample of 20 physicians tested the system, providing feedback via a Likert scale (1-5) on four aspects:
  - Accuracy: alignment with clinician notes
  - Relevance: appropriateness to the query
  - Coverage: completeness of information
  - Clarity: syntax and overall quality

Evaluation Framework - cont'd

G-Eval Evaluation:

- Framework used to evaluate model outputs, consistently using the clinician's notes as a reference.
- Compared against human evaluation.



Automatic Evaluation

Model	Rouge1	Rouge2	RougeL	BLEU	METEOR	Word2Vec	BERTscore
Patient Summary							
GPT4	0.3966	0.1139	0.2380	0.1470	0.2198	0.9035	0.8458
GPT4 32k	0.3963	0.1043	0.2439	0.1742	0.2349	**0.9150**	0.8428
GPT4 Omnni	**0.4286**	**0.1438**	**0.2857**	**0.2013**	**0.2692**	0.8987	**0.8572**
Anthropic	0.3761	0.1073	0.2222	0.1054	0.1905	0.8702	0.8504
List Possible Problems							
GPT4	**0.4444**	**0.1649**	**0.3838**	0.1851	**0.3690**	0.8496	**0.9007**
GPT4 32k	0.4228	0.1322	0.2927	**0.2180**	0.2726	**0.8819**	0.8881
GPT4 Omni	0.3800	0.0612	0.2600	0.1366	0.2709	0.8529	0.8789
Anthropic	0.3579	0.0645	0.2526	0.1221	0.2480	0.8325	0.8810
Email Generation							
GPT4	0.3697	0.0845	0.2129	0.1704	0.2991	0.9164	0.8510
GPT4 32k	0.3371	0.0460	0.1771	0.1153	0.2811	0.8801	0.8443
GPT4 Omni	**0.4536**	**0.1295**	**0.2526**	**0.2345**	**0.3605**	**0.9295**	0.8735
Anthropic	0.4375	0.1166	0.2321	0.2037	0.3037	0.9219	**0.8739**

Human Evaluation

Mean and standard deviation of Human Evaluation scores.

Model	Accuracy	Relevance	Coverage	Clarity
Patient Summary				
GPT4	$\mu = \mathbf{3.83}, \sigma = \mathbf{0.62}$	$\mu = 3.72, \sigma = 0.75$	$\mu = 3.06, \sigma = 0.73$	$\mu = 3.83, \sigma = 0.71$
GPT4 32k	$\mu = 3.72, \sigma = 0.67$	$\mu = \mathbf{3.78}, \sigma = \mathbf{0.81}$	$\mu = 3.22, \sigma = 0.94$	$\mu = \mathbf{4.06}, \sigma = \mathbf{0.54}$
GPT4 Omni	$\mu = 3.83, \sigma = 0.86$	$\mu = 3.50, \sigma = 1.10$	$\mu = 3.33, \sigma = 0.77$	$\mu = 3.89, \sigma = 0.76$
Anthropic	$\mu = 3.83, \sigma = 1.04$	$\mu = 3.83, \sigma = 0.99$	$\mu = \mathbf{3.67}, \sigma = \mathbf{0.59}$	$\mu = 4.17, \sigma = 0.86$
List Possible Problems				
GPT4	$\mu = \mathbf{3.95}, \sigma = \mathbf{0.87}$	$\mu = 4.00, \sigma = 0.84$	$\mu = 3.61, \sigma = 0.85$	$\mu = \mathbf{3.94}, \sigma = \mathbf{0.73}$
GPT4 32k	$\mu = 3.83, \sigma = 0.79$	$\mu = 3.94, \sigma = 0.73$	$\mu = 3.50, \sigma = 0.86$	$\mu = 4.17, \sigma = 0.79$
GPT4 Omni	$\mu = 3.90, \sigma = 1.08$	$\mu = 4.00, \sigma = 0.97$	$\mu = 3.72, \sigma = 1.02$	$\mu = 4.05, \sigma = 0.80$
Anthropic	$\mu = 3.90, \sigma = 0.90$	$\mu = \mathbf{4.06}, \sigma = \mathbf{0.80}$	$\mu = 3.72, \sigma = 1.02$	$\mu = 3.94, \sigma = 0.94$
Email Generation				
GPT4	$\mu = 3.17, \sigma = 0.98$	$\mu = 3.22, \sigma = 0.88$	$\mu = 2.94, \sigma = 0.99$	$\mu = 3.28, \sigma = 0.89$
GPT4 32k	$\mu = 3.00, \sigma = 1.08$	$\mu = 3.56, \sigma = 0.92$	$\mu = 3.33, \sigma = 1.00$	$\mu = 3.44, \sigma = 1.20$
GPT4 Omni	$\mu = 3.78, \sigma = 0.94$	$\mu = \mathbf{4.06}, \sigma = \mathbf{0.64}$	$\mu = \mathbf{3.72}, \sigma = \mathbf{0.67}$	$\mu = 3.94, \sigma = 0.94$
Anthropic	$\mu = 4.06, \sigma = 0.80$	$\mu = 4.11, \sigma = 0.76$	$\mu = 3.90, \sigma = 0.90$	$\mu = 4.06, \sigma = 0.97$

Human Evaluation - cont'd

Inter-rater agreement
for all pairs of raters

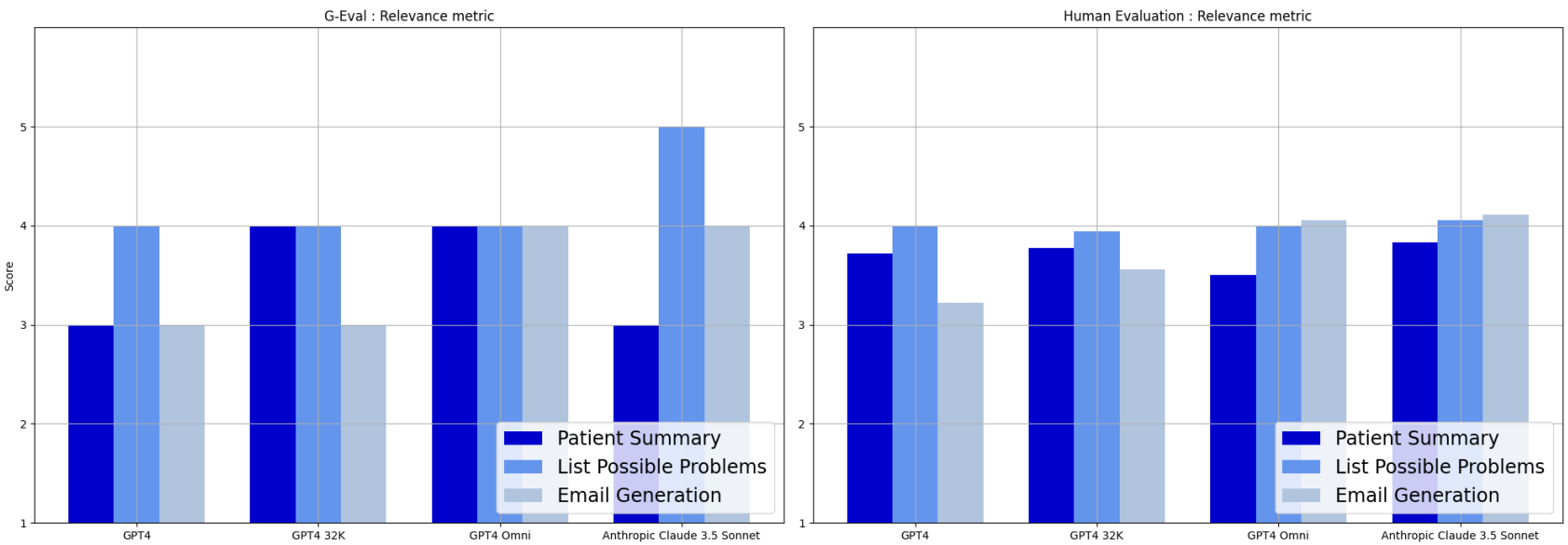
$$P = \{\{R_i = [r_{i,1}, \ldots, r_{i,n}], R_j = [r_{j,1}, \ldots, r_{j,n}]\} | R_i, R_j \in R, R_i \neq R_j\}$$

the corresponding $\kappa_{w_{i,j}}$ values were calculated.
Subsequently, the average $\kappa_w = \frac{1}{|P|} \sum_{(i,j) \in P} \kappa_{w_{i,j}}$ was computed.

GPT4	GPT4 32K	GPT4 Omni	Anthropic
Patient Summary			
$\kappa_w = 0.25$	$\kappa_w = 0.29$	$\kappa_w = 0.37$	$\kappa_w = 0.52$
List Possible Problems			
$\kappa_w = 0.61$	$\kappa_w = 0.63$	$\kappa_w = 0.63$	$\kappa_w = 0.57$
Email Generation			
$\kappa_w = 0.53$	$\kappa_w = 0.61$	$\kappa_w = 0.69$	$\kappa_w = 0.72$

G-Eval Evaluation

Comparison of each metric between G-Eval and Human scores.



Conclusions

Results

- Generative AI can increase productivity
- New proposed framework that includes human and automatic evaluation
  - Sematical metrics work better in this context
  - Raters gave similar scores to shorter text

Conclusions

Results

- Generative AI can increase productivity
- New proposed framework that includes human and automatic evaluation
  - Sematical metrics work better in this context
  - Raters gave similar scores to shorter text

Future Directions

- Conduct a number of measurements with G-Eval equal to the number of raters to compare G-Eval with Human Feedback.