DMIS Lab at ArchEHR-QA 2025: Evidence-Grounded Answer Generation for EHR-based QA via a Multi-Agent Framework

Hyeon Hwang[1*] **Hyeongsoon Hwang**[1*] **Jongmyung Jung**[1] **Jaehoon Yoon**[2,4]
Minju Song[1] **Yein Park**[1] **Dain Kim**[1] **Taewhoo Lee**[1,4] **Jiwoong Sohn**[1,3]
Chanwoong Yoon[1] **Sihyeon Park** [1] **Jiwoo Lee**[1] **Heechul Yang**[1] **Jaewoo Kang**[1,4†]
[1]Korea University [2]Hanyang University [3]ETH Zürich [4]AIGEN Sciences
{hyeon-hwang, hhs8746, kangj}@korea.ac.kr

Abstract

The increasing utilization of patient portals has amplified clinicians' workloads, primarily due to the necessity of addressing detailed patient inquiries related to their health concerns. The ArchEHR-QA 2025 shared task aims to alleviate this burden by automatically generating accurate, evidence-grounded responses to patients' questions based on their Electronic Health Records (EHRs). This paper presents a six-stage multi-agent framework specifically developed to identify essential clinical sentences for answering patient questions, leveraging large language models (LLMs). Our approach begins with OpenAI's o3 model generating focused medical context to guide downstream reasoning. In the subsequent stages, GPT-4.1-based agents assess the relevance of individual sentences, recruit domain experts, and consolidate their judgments to identify essential information for constructing coherent, evidence-grounded responses. Our framework achieved an Overall Factuality score of 62.0 and an Overall Relevance Score of 52.9 on the development set, and corresponding scores of 58.6 and 48.8, respectively, on the test set.

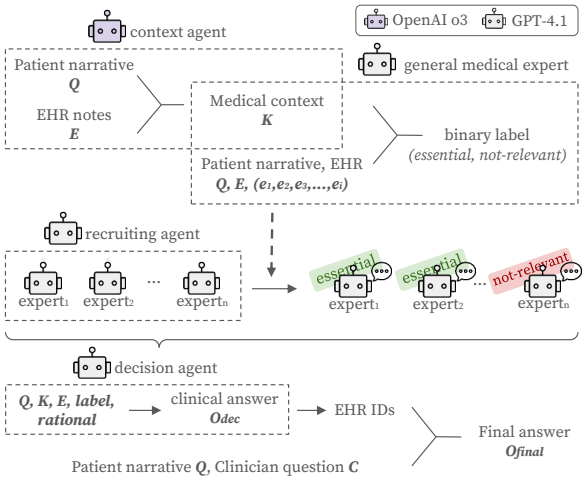

Figure 1: Overview of our six-stage *multi-agent* pipeline for evidence-grounded EHR question answering. The *Context Agent* generates a medical context K from the patient question Q and EHR sentences E. A *General-Medical Expert* labels each sentence as *essential* or *not-relevant* with a brief rationale. The *Recruiting Agent* selects domain-specific *Expert Agents*. The *Decision Agent* integrates all information, generate the answer O_{dec} that cites sentence IDs. The patient narrative, clinician question, and chosen ehr sentences are then assembled into the final reply O_{final}.

1 Introduction

The increased use of patient portals has significantly increased clinicians' workload, especially concerning responding to patients' inbox messages. These messages frequently include detailed questions regarding patients' medical conditions, treatments, and healthcare procedures. Addressing these inquiries manually by clinicians is not only time-consuming but can also delay patient care. To mitigate this burden, the ArchEHR-QA 2025 shared task (Soni and Demner-Fushman, 2025b) focuses on automatically generating accurate and clinically-grounded responses to patients' health-related questions by leveraging information contained within their Electronic Health Records (EHRs).

In this paper, we introduce a six-stage multi-agent framework specifically designed to select appropriate EHR sentences for effectively answering patients' questions in the ArchEHR-QA 2025 shared task. Figure 1 briefly shows how our framework generates evidence-grounded answer based on multiple LLM agents. Our approach begins with context generation using OpenAI's advanced o3 model (OpenAI, 2025b), selected for its demonstrated superior reasoning capabilities in medical contexts. Subsequent stages employ specialized GPT-4.1 (OpenAI, 2025a)-based agents to evaluate the relevance of clinical note sentences individually and collectively, recruit domain-specific

*Equal contribution †Corresponding author

experts dynamically based on the patient's narrative, and integrate diverse expert perspectives into a consensus-driven decision process. The final stage involves synthesizing the identified essential clinical evidence to produce a comprehensive, clinically grounded answer. On the development set, our framework achieved an Overall Factuality score of 62.0 and an Overall Relevance Score of 52.9. On the test set, it attained an Overall Factuality score of 58.6 and an Overall Relevance Score of 48.8.

2 Related Works

Recent advances in large language model (LLM)-based multi-agent systems (Xi et al., 2025; Wang et al., 2023; Guo et al., 2024) have demonstrated significant promise in complex reasoning tasks. Such systems have particularly shown effectiveness in medical domains, with successful implementations like MDAgents (Kim et al., 2024) and MedAgent (Tang et al., 2024) illustrating that deploying specialized agents, each tailored for distinct analytical functions, facilitates robust clinical decision-making processes and precise information extraction. Given the complexity and the sensitive nature of clinical data processing, utilizing a multi-agent framework is particularly well-suited for the ArchEHR-QA 2025 shared task.

3 Method

We propose a six-stage multi-agent framework that automatically extracts the subset of sentences essential for answering patient and clinician questions from a patient's electronic health record (EHR) corpus. The context generation stage employs OpenAI o3 (OpenAI, 2025b) selected for its superior reasoning performance to generate focused and clinically relevant medical context, whereas the remaining stages rely on specialised agents based on GPT-4.1 (OpenAI, 2025a) that provide complementary analytic perspectives and collectively converge on a final consensus.

3.1 Problem Definition

Throughout this section, Q denotes the patient narrative, $E = \{e_1, \ldots, e_n\}$ the set of EHR sentences, and $S \subseteq E$ the subset labeled essential. An agent is defined as a function of the form:

$$A_{\text{role}}(\text{M}, \text{I}) = \text{O},$$

- ***model*** **M** is an instantiated LLM (e.g. OpenAI o3 or GPT-4.1).

ContextAgent Example

```
### Patient Question
ICU 15 days for severe abdominal pain; diagnosed with
common-bile-duct (CBD) sludge and started Udiliv,
but doctor still advises ERCP.
Can medication alone clear the sludge?

### Generated Medical Context
1. UDCA may dissolve microscopic gallbladder sludge
but not obstructive CBD sludge, especially
when infection or jaundice is present...

2. Despite of ICU care and ongoing Udiliv,
the sludge has persisted strong enough evidence
that medication has not yet relieved the obstruction.
...
```

Figure 2: An example of medical contexts generated by the *ContextAgent*, with long explanations truncated for brevity.

- ***input*** **I** is a role-specific set of inputs, comprising prompts, auxiliary context, and intermediate metadata.
- ***output*** **O** is the structured result expected from that role (e.g. a context paragraph, a binary relevance label with rationale.)

3.2 Multi-Agent Framework

Context Generation. The context agent A_{ctx} uses the high-performance model M_{ctx} (OpenAI o3) to generate a medical context K to address the patient question Q, guiding downstream reasoning.

$$K = A_{\text{ctx}}\big(M_{\text{ctx}},\ (Q, E)\big)$$

General Medical Expert Relevance Screening. The general medical expert agent A_{gen} evaluates the essentiality of each individual EHR sentence. The agent is provided with the patient's question Q, the generated medical context K, the full set of EHR notes E, and the specific sentence e_i under evaluation. The agent outputs a binary label $\ell_i \in \{\text{essential}, \text{not-relevant}\}$, indicating the essentiality of e_i, along with a rationale r_i. This process is repeated independently for each sentence.

$$(\ell_i^{\text{gen}}, r_i^{\text{gen}}) = A_{\text{gen}}\big(M_{\text{gen}},\ (Q, K, E, e_i)\big)$$

Experts Recruitment. The recruiting agent A_{rec} synthesizes the patient narrative Q, medical context K, and the full set of EHR notes E to assemble an expert panel, denoted as

$$\text{Experts} = \{A_{\text{exp}}^{(1)}, \ldots, A_{\text{exp}}^{(m)}\} = A_{\text{rec}}(Q, K, E).$$

Domain-Specific Assessment. Each expert agent $A_{\text{exp}}^{(j)} \in \text{Experts}$ then receives the (Q, K, E) to perform a domain-specific evaluation. Based on this comprehensive input, the expert evaluates the essentiality of all sentences in the EHR collectively, leveraging their specialized medical knowledge to make sentence-level judgments. The relevance label and rationale set produced by the j-th expert agent are given by

$$(\boldsymbol{L}^{(j)}, \boldsymbol{R}^{(j)}) = A_{\text{exp}}^{(j)}\big(M_{\text{exp}}^{(j)}, (Q, K, E)\big),$$

where $\boldsymbol{L}^{(j)} = (\ell_1^{(j)}, \ldots, \ell_n^{(j)})$ represents the set of sentence-level labels, and $\boldsymbol{R}^{(j)} = (r_1^{(j)}, \ldots, r_n^{(j)})$ represents the corresponding rationales provided by the expert for each sentence $e_i \in E$. Each label $\ell_i^{(j)} \in \{\text{essential}, \text{not-relevant}\}$ encodes the expert's judgment regarding the essentiality of sentence e_i, and each rationale $r_i^{(j)}$ provides the justification for that judgment.

Consensus Integration. The aggregated package

$$I_{\text{dec}} = \Big(Q,\ K,\ E,\ (\ell_i^{\text{gen}}, r_i^{\text{gen}})_{i=1}^{n},\ \{(\mathbf{L}^{(j)}, \mathbf{R}^{(j)})\}_{j=1}^{m}\Big)$$

is forwarded to the decision agent A_{dec}. This agent consolidates the upstream judgments to determine the definitive essential-sentence set S, and uses S to craft a comprehensive, evidence-grounded clinical answer O_{dec} to the patient's question Q.

$$O_{dec} = A_{\text{dec}}\big(M_{\text{dec}}, I_{\text{dec}}\big),$$

Final Answer Generation. Finally, we extract the IDs of the essential notes identified in the O_{dec}, and then concatenate the patient narrative, clinician question, and the selected essential notes to generate a comprehensive response O_{final}.

4 Experimental settings

4.1 Dataset

To evaluate our framework, we utilize the benchmark dataset (Soni and Demner-Fushman, 2025a) provided by ArchEHR-QA 2025. This dataset consists of case-based collections, each comprising a patient narrative, a patient question, a clinician question, and associated EHR data intended to support answering the question. The EHR data for each case is composed of multiple sentences, each annotated with a unique sentence ID. The dataset consists of a development set and a test set. Among these, only the development set provides sentence-level relevance labels (categorized as essential, supplementary, or not relevant) for evaluating the performance of answer generation.

Method	Factuality (Strict Micro) Precision	Recall	F1
Multiclass classification			
w/ Experts	**64.8**	52.1	57.8
w/ Context K	61.0	52.1	56.2
w/ Experts + Context K	64.2	52.1	57.6
Binary classification			
w/ Experts	50.0	69.5	58.1
w/ Context K	52.1	69.5	59.6
w/ Experts + Context K	53.4	**73.9**	**62.0**

Table 1: Factuality score comparison for multiclass models (*essential* / *supplementary* / *not-relevant*) and binary models (*essential* / *not-relevant*) using *w/ Experts*, *w/ Context K*, and *w/ Experts + Context K*.

4.2 Metrics

We adopt three evaluation metrics in accordance with the official scoring criteria of ArchEHR-QA 2025: *Overall Factuality Score*, *Overall Relevance Score*, and *Overall Score*.

Overall Factuality Score measures the F1 score between the set of sentence IDs cited in the final answer and those cited in the gold answer. This score is computed based on the counts of true positives, false positives, and false negatives aggregated across each case.

Overall Relevance Score evaluates the semantic and lexical similarity between the final and gold answers using a combination of BLUE (Papineni et al., 2002), ROUGE (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang* et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023) metrics. The final score is obtained by combining the normalized scores of these individual metrics. The gold answer in this context is constructed by concatenating the patient narrative, clinician question, and essential EHR sentences provided for each case.

Overall Score serves as the primary evaluation metric for this challenge. It is defined as the average of the *Overall Factuality Score* and the *Overall Relevance Score*.

4.3 Results

To verify the quality of our framework, this section presents (i) a sentence-level factuality analysis on the development set, (ii) a multi-metric relevance

Method	Relevance						
	BLEU	ROUGE-L	SARI	B.S.	A.S.	MEDCON	Overall
Questions + Predicted sentences	**19.2**	**53.6**	34.8	**58.2**	**97.6**	**54.1**	**52.9**
Summary of predicted sentences	2.0	25.1	**55.7**	27.6	42.6	37.4	31.7

Table 2: Comparison of relevance scores between two answer generation methods: using full question with predicted essential sentences vs. using a summary of the predicted essential sentences. Abbreviations: ROUGE-L= ROUGE-Lsum, B.S. = BERTScore, A.S. = AlignScore.

Team	Overall	Factuality (Strict Micro)			Relevance						
		Precision	Recall	F1	BLEU	ROUGE-L	SARI	B.S.	A.S.	MEDCON	Overall
DMIS Lab (Ours)	**53.7**	57.9	59.3	58.6	**14.3**	**46.5**	36.7	**53.9**	**92.4**	**49.3**	**48.8**
Neural	51.5	55.4	63.8	59.3	8.5	34.1	**73.1**	39.1	67.3	40.0	43.7
LAILab	51.0	56.0	**65.5**	**60.4**	6.5	32.7	69.2	37.4	65.3	38.4	41.6
LAMAR	49.1	60.6	53.6	56.9	6.0	32.1	65.8	36.4	64.3	43.6	41.4
ssagarwal	45.0	**68.8**	36.2	47.5	4.7	31.1	70.0	36.9	74.9	38.0	42.6

Table 3: Official results of the leaderboard (Top 5) on ArchEHR-QA 2025 dataset. The teams are ranked based on Overall score. Abbreviations: ROUGE-L = ROUGE-Lsum, B.S. = BERTScore, A.S. = AlignScore.

analysis, and (iii) a comparison of test-set scores on the official ArchEHR-QA 2025 leaderboard.

4.3.1 Factuality analysis

The sentence–level evaluation, summarised in Table 1, reveals a clear benefit from contextual conditioning. The *multiclass* (essential / supplementary / not-relevant) variant achieves the highest precision (62.2%) but simultaneously records the lowest recall (51.4%), resulting in an F1 of 56.3. Conversely, the *binary* (essential / not-relevant) classifier attains the greatest recall (69.5%) at the expense of precision (50.0%), yielding an F1 of 58.1. When the identical binary classification approach is prefixed with the automatically generated medical context K, recall increases further to 73.9% while precision recovers to 53.4%, producing the best strict-micro F1 of **62.0**. These results indicate that (i) finer-grained labels do not compensate for the recall penalty inherent in multiclass formulations, and (ii) domain-aware context provides the disambiguating cues necessary to recover clinically critical sentences, thereby maximising overall factuality.

4.3.2 Relevance analysis

Table 2 compares two answer-construction strategies. Passing the generator the *question* concatenated with the sentences predicted *essential* achieves higher scores across most relevance metrics: BLEU increases from 2.0 to 19.2, ROUGE-Lsum from 25.1 to 53.6, AlignScore from 42.6 to 97.6, and MEDCON from 37.4 to 54.1, resulting in an overall relevance score of **52.9**. By contrast, generating a free-form *summarised answer* results in an overall relevance score of 31.7. Based on this result, we adopted the *questions + essential sentences* strategy for our final test submission.

4.3.3 Official Leaderboard

Table 3 summarises official test-set results. Our system (DMIS Lab) ranks first with an Overall score of 53.7, balancing a factuality F1 of 58.6 and a relevance overall of 48.8. The consistency between development and test splits underscores the effectiveness of the proposed multi-agent architecture.

5 Conclusion

In this paper, we presented a multi-agent framework for answering patients' health-related questions using their EHRs. Our method decomposes the task into distinct stages: context generation, relevance assessment, expert recruitment, and consensus integration. Each stage is handled by specialized LLM-based agents. This structured, modular approach enables robust identification of essential clinical sentences and the generation of coherent, evidence-grounded responses. Our framework achieved strong performance on both the development and test sets in terms of factuality. These results highlight the potential of LLM-based multi-agent systems in clinical question answering and suggest promising directions for future work in automating patient-clinician communication based on real EHR data.

Acknowledgements

This research was supported by (1) the National Research Foundation of Korea (NRF-2023R1A2C3004176, RS-2023-00262002), (2) the Ministry of Health & Welfare, Republic of Korea (HR20C002103), and (3) ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201819). This work was supported by the Korea Bio Data Station(K-BDS) with computing resources including technical support.

Limitations

De-identification Assumptions. ArchEHR-QA 2025 provides de-identified notes, but real clinical systems often contain partially identifiable information. Our framework does not include additional privacy-preserving mechanisms and would need adaptation before deployment on raw, identifiable EHR data.

Dependence on Closed source LLMs. Our framework relies on OpenAI's o3 and GPT-4.1 models. Although these models currently provide state-of-the-art reasoning, they are proprietary, incur non-trivial inference costs, and can change without notice. Reproducing or extending our results with fully open-source alternatives may require prompt and hyper-parameter retuning.

Latency and Cost. The framework's inference time and computational cost remain substantial, posing challenges for real-time deployment in high-volume patient-portal environments. These resource demands may limit its practical scalability without further optimization.

References

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2025a. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-05-09.

OpenAI. 2025b. o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Accessed May 9, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset

for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Prompt example for Each Agent

ContextAgent system_message

```
- You are a ContextBuildingAgent.

- Your role is to carefully read the patient's question and generate a relevant medical context
    that would help answer it accurately.

Guidelines:
- Focus only on medically meaningful information that would assist in answering the patient's
    question.
- Include important background knowledge, clinical reasoning steps, diagnostic considerations,
    and treatment options that are directly relevant to the question.
- Do not fabricate information unrelated to the patient's case.
- The context should be clear, concise, medically accurate, and structured to support clinical
    decision-making.

Output:
- Plain text only.
- Write in a factual and professional tone as if you are preparing supporting information for a
    medical expert.
```

RecruitingAgent prompt

```
- You are a recruiting agent.

- Given a patient question and EHR note, your task is to identify the most relevant medical
    experts.

- Return a JSON object with a key called 'experts' whose value is a list of strings.
- Example: {"experts": ["cardiology", "gastroenterology"]}

- Do not include any explanation or additional text. Only return the JSON object.
```

AnalysisAgent prompt

```
- You are a medical reviewer. Your task is to evaluate whether each individual sentence
in a clinical note is relevant to answering the patient's question.

- Each sentence is identified by its ID. For every sentence, return:
Whether the sentence is 'essential' or 'not-relevant', and A brief justification for your
judgment, explaining why the sentence does or does not contribute to answering the question.

- A sentence is considered **essential** if it directly or indirectly helps answer the question
through evidence, explanation, clarification, or medically meaningful context.
- Avoid marking sentences as essential if they only provide background or loosely related
    information.
```

ExpertAgent prompt

```
- You are a board-certified clinical expert in {expertise}.

- You are evaluating each sentence in an EHR note from the unique clinical
perspective of your own specialty ({expertise}).

- Your role is to assess whether the sentence meaningfully contributes to
answering the patient's question, based on your specialty's reasoning
principles, typical clinical decision-making, and domain-specific interpretation.

- If the sentence contains medically meaningful evidence, logic, or interpretation
```

```
  that a {expertise} specialist would find critical to answer the question,
  label it "ESSENTIAL".

- If the sentence contains no valuable insight or decision-making relevance from
  your specialty perspective, label it "NOT-RELEVANT".

- Avoid generic reasoning. Always ground your decision in your expert role.
```

DecisionAgent prompt

```
- You are a skilled medical expert. Your task is to provide an accurate and evidence-based answer
     to a patient's question using the provided EHR note.

- Your answer must be medically sound and supported by evidence extracted from the provided EHR
    note sentences.

- When composing your answer, you **must** include citation IDs (enclosed in pipe symbols |, for
    example, |3,4|) only for the parts of your answer that are directly supported by evidence
    from the EHR note.

- Each sentence in your answer should be on a separate line.

- **Before writing your answer, carefully verify whether the EHR note includes any sentences that
     are truly relevant to answering the patient's question.**
```

A.2 Basic Structure of `Agent`

Listing A.2: `Agent` Class Definition

```python
class Agent:
    def __init__(self, agent_name,
                 model='model',
                 temperature=0,
                 system_message='You are a helpful assistant.'):
        self.agent_name  = agent_name
        self.model       = model
        self.temperature = temperature
        self.system_message = system_message
        self.client = openai.OpenAI()

    def generate_response(self, user_msg: str) -> str:
        rsp = self.client.chat.completions.create(
            model=self.model,
            messages=[{"role": "system",
                       "content": self.system_message},
                      {"role": "user",
                       "content": user_msg}],
            temperature=self.temperature,
            max_tokens=2048,
        )
        return rsp.choices[0].message.content.strip()
```