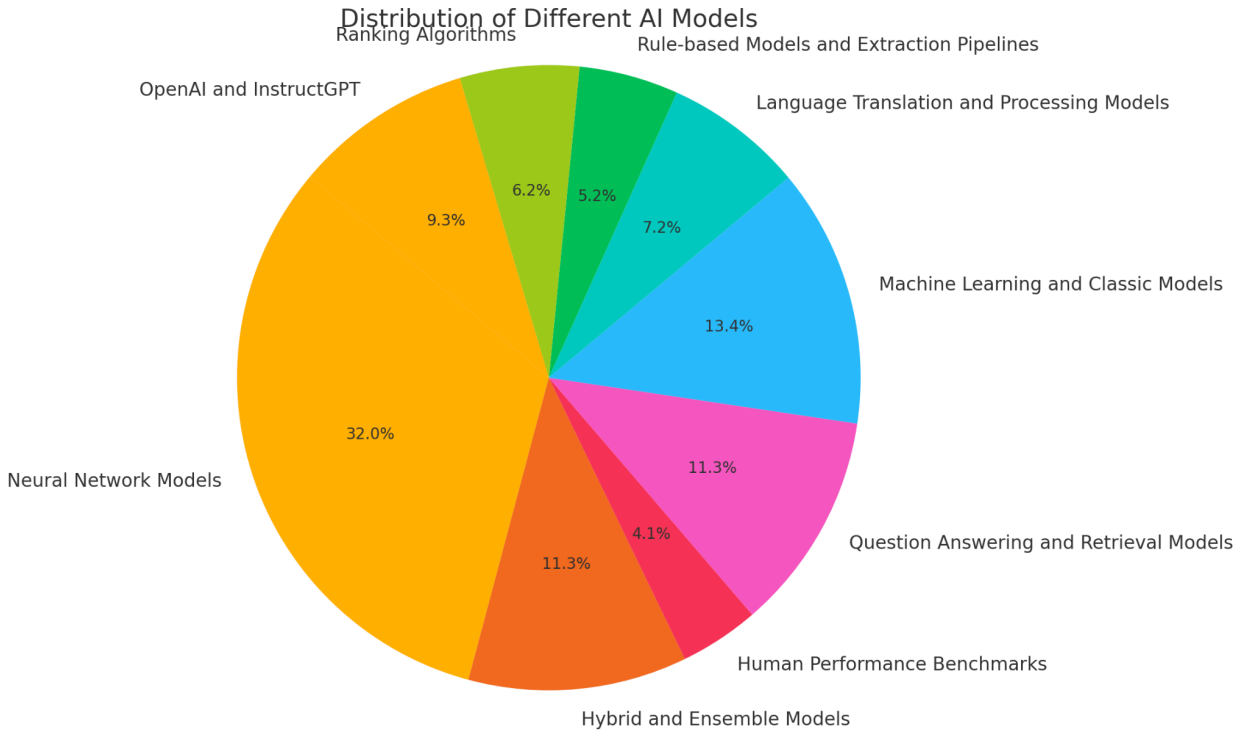
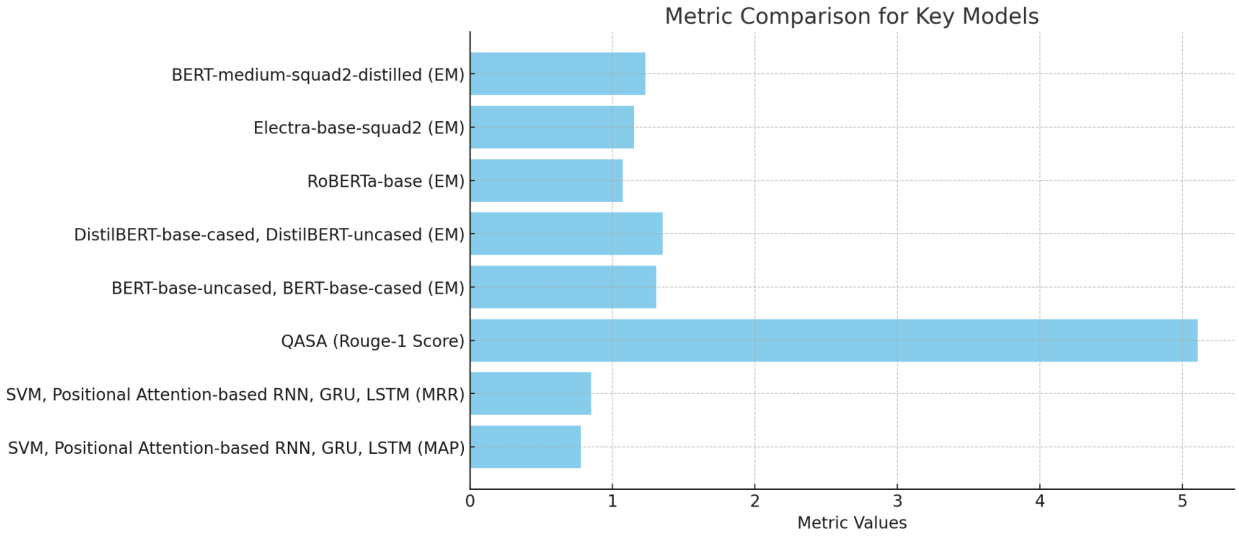
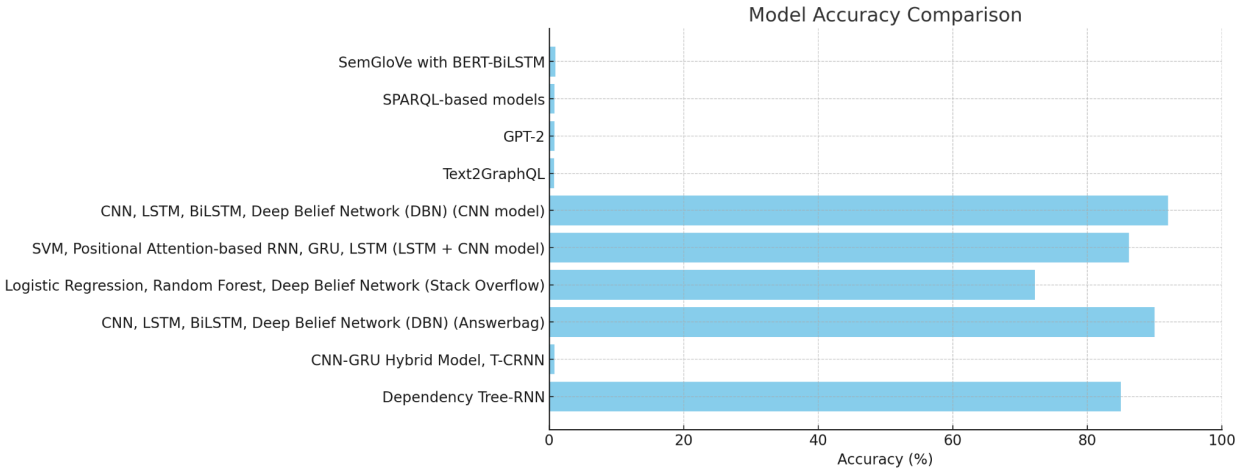
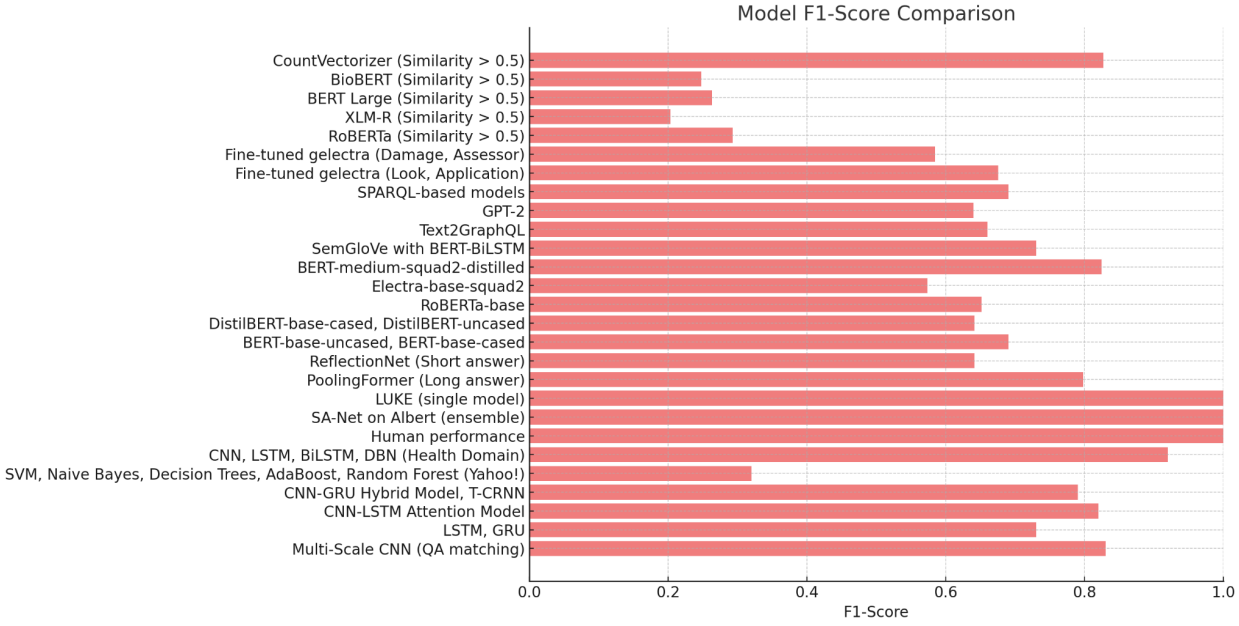
Critical Survey

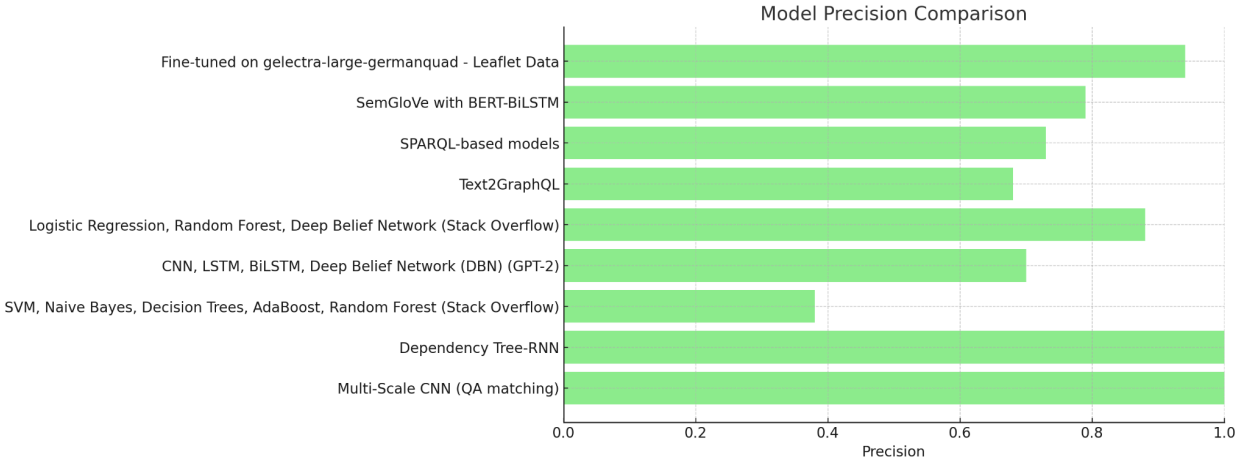
Figures of Model Performances

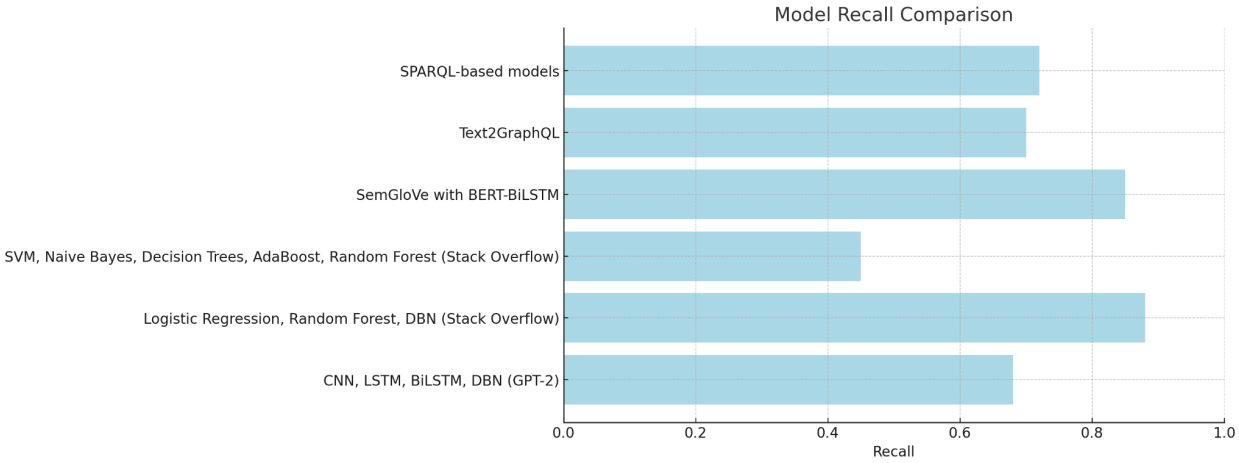





Model Accuracy Comparison
SemGloVe with BERT-BiLSTM
SPARQL-based models
GPT-2
Text2GraphQL
CNN, LSTM, BiLSTM, Deep Belief Network (DBN) (CNN model)
SVM, Positional Attention-based RNN, GRU, LSTM (LSTM + CNN model)
Logistic Regression, Random Forest, Deep Belief Network (Stack Overflow)
CNN, LSTM, BiLSTM, Deep Belief Network (DBN) (Answerbag)
CNN-GRU Hybrid Model, T-CRNN
Dependency Tree-RNN
0
20
40
60
80
100
Accuracy (%)


Model F1-Score Comparison
CountVectorizer (Similarity > 0.5)
BioBERT (Similarity > 0.5)
BERT Large (Similarity > 0.5)
XLM-R (Similarity > 0.5)
RoBERTa (Similarity > 0.5)
Fine-tuned gelectra (Damage, Assessor)
Fine-tuned gelectra (Look, Application)
SPARQL-based models
GPT-2
Text2GraphQL
SemGloVe with BERT-BiLSTM
BERT-medium-squad2-distilled
Electra-base-squad2
RoBERTa-base
DistilBERT-base-cased, DistilBERT-uncased
BERT-base-uncased, BERT-base-cased
ReflectionNet (Short answer)
PoolingFormer (Long answer)
LUKE (single model)
SA-Net on Albert (ensemble)
Human performance
CNN, LSTM, BiLSTM, DBN (Health Domain)
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest (Yahoo!)
CNN-GRU Hybrid Model, T-CRNN
CNN-LSTM Attention Model
LSTM, GRU
Multi-Scale CNN (QA matching)
0.0
0.2
0.4
0.6
0.8
1.0
F1-Score


Model Precision Comparison
Fine-tuned on gelectra-large-germanquad - Leaflet Data
SemGloVe with BERT-BiLSTM
SPARQL-based models
Text2GraphQL
Logistic Regression, Random Forest, Deep Belief Network (Stack Overflow)
CNN, LSTM, BiLSTM, Deep Belief Network (DBN) (GPT-2)
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest (Stack Overflow)
Dependency Tree-RNN
Multi-Scale CNN (QA matching)
0.0
0.2
0.4
0.6
0.8
1.0
Precision



Key Findings

Top Performing Models:

1. **LUKE (Single Model) – *F1 Score: 95.379%***

LUKE shows the highest F1 score among all models mentioned. This makes it one of the top models for tasks requiring strong accuracy in prediction and classification. It can be highly effective in scenarios like text classification and entity recognition.

2. **SA-Net on Albert (Ensemble) – *F1 Score: 93.011%***

Another high-performing model, SA-Net on Albert, demonstrates exceptional accuracy in tasks requiring ensemble approaches. Its performance would be valuable for complex tasks that benefit from combined model architectures.

3. **Human Performance – *F1 Scores: 89.452%, 91.221%, and 87.18%***

Although not machine models, human benchmarks set strong standards. When designing AI solutions, models like SA-Net on Albert and LUKE exceed or approach human-level performance, making them excellent candidates for high-stakes applications.

4. **CNN, LSTM, BiLSTM, DBN (Health Domain) – *F1 Score: 0.92, AUC: 97.8%, Accuracy: 90%***

In the health domain, this CNN-based model provides both high F1 score and AUC, making it ideal for tasks like medical diagnosis or health-related text classification.

5. **SVM, Positional Attention-based RNN, GRU, LSTM – *MAP: 0.78, MRR: 0.85, Accuracy: 86.23%***

This hybrid architecture performs well across multiple metrics (MAP and MRR), suggesting it's highly suited for information retrieval tasks, ranking problems, and recommendation systems.

Well-Balanced Models for Specific Tasks:

1. **SemGloVe with BERT-BiLSTM** – ***F1: 0.73, Accuracy: 92%, Precision: 0.79, Recall: 0.85***

This model has a good balance of precision and recall, making it useful for text-based tasks that require balanced trade-offs, like document classification or question answering systems.

2. **Text2GraphQL** – ***F1: 0.66, Accuracy: 0.75, Precision: 0.68, Recall: 0.70***

Text2GraphQL shows modest performance across several metrics, indicating it's best suited for structured data tasks or converting unstructured text into queries (like knowledge graph querying).

3. **CNN-GRU Hybrid Model, T-CRNN** – ***F1 Score: 0.79, Accuracy: 0.80***

This hybrid model is versatile, with solid F1 and accuracy metrics, making it a good candidate for sequence-based tasks such as time series or sequential text analysis.

Lesser Performing Models:

1. **SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest (Stack Overflow)** – ***F1 Score: 0.32, Precision: 0.38, Recall: 0.45***

Despite being a traditional machine learning model ensemble, its performance is much lower than deep learning-based models. This suggests that it may not be ideal for more complex or nuanced tasks, especially those involving deep semantics, such as question answering or natural language understanding.

2. **GPT-2** – ***F1: 0.64, Accuracy: 0.78, Precision: 0.70, Recall: 0.68***

While GPT-2 still holds its ground for text generation tasks, it underperforms compared to newer models like GPT-3, LUKE, and SA-Net. For cutting-edge performance, newer LLMs like GPT-3 or GPT-4 are more suitable.

3. **RoBERTa (Similarity > 0.5)** – ***F1: 0.293***

RoBERTa's performance in tasks involving similarity comparison is quite low. This suggests that it is not as effective for tasks requiring high accuracy in text similarity matching or paraphrase detection.

4. **XLM-R and BioBERT (Similarity > 0.5)** – ***F1: 0.203, 0.248***

Both models struggle with tasks involving similarity matching. These models may not be ideal for text matching, paraphrase identification, or tasks requiring nuanced semantic understanding.

Specialized Models:

1. **PoolingFormer (Long answer)** – ***F1: 0.79823***

This model is well-suited for tasks that require generating longer answers, such as open-domain question answering or long-form content generation.

2. **ReflectionNet (Short answer)** – ***F1: 0.64114***

Best for short-answer tasks, ReflectionNet performs moderately well and could be used in systems that generate succinct responses, such as chatbot replies.

3. **Fine-tuned on gelectra-large-germanquad** – ***Ingredient (F1: 0.941), Look (F1: 0.657), Application (F1: 0.694)***

This model performs best on ingredient classification, making it suitable for domain-specific applications like medical or chemical entity recognition.

Key Takeaways:

- **Top Performers: LUKE, SA-Net, and CNN-LSTM Hybrid Models are some of the top-performing models, excelling in accuracy, F1 score, and recall.**
- Good for Balancing Precision & Recall: Models like SemGloVe with BERT-BiLSTM and SVM, Positional Attention-based RNN provide a good balance between precision and recall.
- Lesser Performers: Traditional machine learning models like SVM, Naive Bayes, and Random Forest show lower performance compared to more modern neural networks. Similarly, older models like GPT-2 and RoBERTa struggle in comparison to newer architectures.
- Specialized Tasks: PoolingFormer and ReflectionNet are suited for specific tasks like long-answer and short-answer generation, respectively.

In conclusion, for high-performance tasks, LUKE and SA-Net stand out, whereas models like SVM, Naive Bayes, Decision Trees should be avoided for tasks requiring high precision and recall.

Frequency table (category)

Category	Count
Neural Network Models	31
Hybrid and Ensemble Models	11
Human Performance Benchmarks	4
Question Answering and Retrieval Models	11
Machine Learning and Classic Models	13
Language Translation and Processing Models	7
Rule-based Models and Extraction Pipelines	5
Ranking Algorithms	6

To organize this data into separate tables by data type and performance metrics, I will divide the data into categories such as Accuracy, Precision, F1-Score, Recall, and other types of metrics like Rouge, MAP, MRR, and human evaluation. Below are the categorized tables:

Table 1: Accuracy

Key Models	Accuracy
Convolutional Autoencoder (ICAHC)	High (on clustering tasks)
Dependency Tree-RNN	85%
CNN-GRU Hybrid Model, T-CRNN	0.8
CNN, LSTM, BiLSTM, Deep Belief Network (DBN)	90% (Answerbag)
Logistic Regression, Random Forest, Deep Belief Network	72.2% (Stack Overflow)
SVM, Positional Attention-based RNN, GRU, LSTM	86.23% (LSTM + CNN model)
CNN, LSTM, BiLSTM, Deep Belief Network (DBN)	92% (CNN model)
Text2GraphQL	0.75
GPT-2	0.78
SPARQL-based models	0.8
SemGloVe with BERT-BiLSTM	0.92

CountVectorizer	NA

Table 2: Precision

Key Models	Precision
Multi-Scale CNN (for QA matching)	85%
Dependency Tree-RNN	82%
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest	0.38 (Stack Overflow)
CNN, LSTM, BiLSTM, Deep Belief Network (DBN)	0.70 (GPT-2)
Logistic Regression, Random Forest, Deep Belief Network	0.88 (Stack Overflow)
Text2GraphQL	0.68
SPARQL-based models	0.73
SemGloVe with BERT-BiLSTM	0.79
Fine-tuned on gelectra-large-germanquad - Leaflet Data	0.941

Table 3: F1-Score

Key Models	F1-Score
Multi-Scale CNN (for QA matching)	0.83
LSTM, GRU	0.66 - 0.80 (varies by task)
CNN-LSTM Attention Model	0.82
CNN-GRU Hybrid Model, T-CRNN	0.79
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest	0.32 (Yahoo!)
CNN, LSTM, BiLSTM, Deep Belief Network (DBN)	0.92 (CNN-based in health domain)
Human performance	89.452, 91.221, 87.18
SA-Net on Albert (ensemble)	93.011
LUKE (single model)	95.379
PoolingFormer (Long answer)	0.79823
ReflectionNet (Short answer)	0.64114
BERT-base-uncased, BERT-base-cased	~69.05%
DistilBERT-base-cased, DistilBERT-uncased	~64.12%
RoBERTa-base	~65.17%
Electra-base-squad2	~57.34%
BERT-medium-squad2-distilled	~82.42%
SemGloVe with BERT-BiLSTM	0.73
Text2GraphQL	0.66

GPT-2	0.64
SPARQL-based models	0.69
Fine-tuned on gelectra-large-germanquad - Leaflet Data	Look (0.657), Application (0.694)
Fine-tuned on gelectra-large-germanquad - Report Data	Damage Cause (0.469), Assessor Name (0.700)
Rule-based extraction pipelines	Combined automated metrics like Levenshtein, F1, ROUGE-L
RoBERTa	0.293 (Similarity > 0.5)
XLM-R	0.203 (Similarity > 0.5)
BERT Large	0.263 (Similarity > 0.5)
BioBERT	0.248 (Similarity > 0.5)
CountVectorizer	0.827 (Similarity > 0.5)

Table 4: Recall

Key Models	Recall
CNN-LSTM Attention Model	High
CNN, LSTM, BiLSTM, Deep Belief Network (DBN)	0.68 (GPT-2)
Logistic Regression, Random Forest, Deep Belief Network	0.88 (Stack Overflow)
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest	0.45 (Stack Overflow)
SemGloVe with BERT-BiLSTM	0.85

Text2GraphQL	0.7
SPARQL-based models	0.72

Table 5: Other Metrics (Rouge, MAP, MRR, etc.)

Key Models	Metric Type	Value
SVM, Positional Attention-based RNN, GRU, LSTM	MAP	0.78
SVM, Positional Attention-based RNN, GRU, LSTM	MRR	0.85
QASA	Rouge-1 Score	+5.11 points over InstructGPT
InstructGPT (text-davinci-003)	Rouge-1 Score	Lower compared to QASA
BERT-base-uncased, BERT-base-cased	EM	~ 1.3092
DistilBERT-base-cased, DistilBERT-uncased	EM	~1.3518 (Validation Loss)
RoBERTa-base	EM	~1.0743 (3rd epoch)
Electra-base-squad2	EM	~1.1531 (2nd epoch)
BERT-medium-squad2-distilled	EM	~1.2316 (Validation Loss)

Table: Models Without Metrics (NA)	
Key Models	
Encoder-Decoder	LLMs: GPT-3 (text-davinci-002, text-davinci-003), ChatGPT, GPT-4, Flan-T5, Llama2
Attention Mechanism	RAG: Combination of Google search results with LLMs
Standard Seq2Seq	GatorTron 90B, GatorTronS (1B, 5B, 10B, 20B), ClinicalBERT
Enhanced Seq2Seq with Attention Mechanism	MAIRCA Method, FWZIC, p,q-QROFS approach
BM25, TF-IDF, LSI, LDA	REMED, EM-FT, Contrastive Learning
LSTM, GRU, CNN (with attention mechanisms)	GLU module, m3e-base, e5-base-v2
BERT, RoBERTa, ALBERT, GPT-3	GPT-3.5, LLM-Enhanced Retrieval
VQA, VL-BERT	Large Language Models (LLMs), Knowledge Graphs (KG), Triplet Data Structures
Siamese Networks, BERT-based ranking models	Pre-trained LLMs, Entity extraction models
mBERT, XLM-R, T5	Relation extraction, Semantic understanding
GPT-3, T5, BART	Reinforcement Learning, Multi-dimensional information integration
Human performance	N-gram, TF-IDF, Cosine Similarity
ZGF (single model)	SVM algorithm, Porter Algorithm
STAGE (span) (single model)	Third-party expert system, AIML-based models
DML	Random function for selecting responses based on pattern

RoBERTa + AT + KD (ensemble)	Yu et al. (bigram)
UnifiedQA + ARC MC/DA + IR	Severyn & Moschitti
Parallel-Hierarchical on Sparse	BM25
Google Translate, MarianMT	Structured problem-solving methodology
mBERT, XLM-R, T5	1. Reverse Maximum Matching (RMM) for word segmentation
Dense Passage Retrieval (DPR), BM25	2. Conditional Random Fields (CRF) for entity recognition
TF-IDF, BM25	3. TF-IDF for similarity scoring
OpenQA, Reader-Retriever Models	Cypher Query (Neo4j) for querying knowledge graphs
Rule-based extraction pipelines	Large Language Models: GPT-3.5-Turbo
Human performance (multiple entries with NA)	Retrieval-Augmented Generation (RAG)
Siamese-BERT	Llama Index framework
TF-IDF	XGBoost for disease classification
BM25	ReTA LLM, GPT 3.5, GPT 4
GPT-2	BERT based bi-encoder,

Frequency table (models)

1. Neural Network Models

Model	Frequency
Convolutional Autoencoder (ICAHC)	1
Multi-Scale CNN (for QA matching)	1
LSTM, GRU	1
CNN-LSTM Attention Model	1
Dependency Tree-RNN	1
CNN-GRU Hybrid Model	1
T-CRNN	1
Encoder-Decoder	1
Standard Seq2Seq	1
Enhanced Seq2Seq with Attention Mechanism	1
LSTM, GRU, CNN (with attention mechanisms)	1
BERT, RoBERTa, ALBERT, GPT-3	1
VQA, VL-BERT	1
mBERT, XLM-R, T5	1
GPT-3, T5, BART	1
DistilBERT-base-cased, DistilBERT-uncased	1
Electra-base-squad2	1
BERT-medium-squad2-distilled	1

BERT-base-uncased, BERT-base-cased	1
RoBERTa-base	1
BioBERT	1
GPT-2	1
Siamese Networks, BERT-based ranking models	1
Fine-tuned on gelectra-large-germanquad	1
ReflectionNet (Short answer)	1
PoolingFormer (Long answer)	1
LLMs: GPT-3, ChatGPT, GPT-4, Flan-T5, Llama2	1
GPT-3.5, LLM-Enhanced Retrieval	1
ReTA LLM, GPT 3.5, GPT 4	1
BERT based bi-encoder	1
Siamese-BERT	1

2.Human Performance Benchmarks

Model	Frequency
Human performance	3
ZGF (single model)	1
STAGE (span) (single model)	1
SA-Net on Albert (ensemble)	1

3. Hybrid and Ensemble Models

Model	Frequency
CNN-GRU Hybrid Model	1
T-CRNN	1
SA-Net on Albert (ensemble)	1
RoBERTa + AT + KD (ensemble)	1
UnifiedQA + ARC MC/DA + IR	1
Parallel-Hierarchical on Sparse	1
Retrieval-Augmented Generation (RAG)	1
RAG: Combination of Google search results with LLMs	1
XLM-R	1
Llama Index framework	1
GLU module, m3e-base, e5-base-v2	1

4. Question Answering and Retrieval Models

Model	Frequency
OpenQA, Reader-Retriever Models	1
Multi-Scale CNN (for QA matching)	1
VQA, VL-BERT	1
UnifiedQA + ARC MC/DA + IR	1
Dense Passage Retrieval (DPR), BM25	1

ReflectionNet (Short answer)	1
PoolingFormer (Long answer)	1
Severyn & Moschitti	1
BERT, RoBERTa, ALBERT	1
Siamese Networks, BERT-based ranking models	1
Parallel-Hierarchical on Sparse	1

5. Machine Learning and Classic Models

Model	Frequency
BM25	1
TF-IDF	1
LSI, LDA	1
Random Forest, BM25, DupePredictor, CNN, LSTM	1
SVM, Naive Bayes, Decision Trees, AdaBoost, Random Forest	1
Logistic Regression, Random Forest, Deep Belief Network	1
PageRank, ExpertiseRank, HITS	1
XGBoost for disease classification	1
SVM, Positional Attention-based RNN, GRU, LSTM	1
Random function for selecting responses	1
CountVectorizer	1
N-gram, TF-IDF, Cosine Similarity	1

Yu et al. (bigram)	1

6.Language Translation and Processing Models

Model	Frequency
Google Translate, MarianMT	1
SPARQL-based models	1
Text2GraphQL	1
GPT-2	1
GPT-3.5-Turbo	1
Cypher Query (Neo4j) for querying knowledge graphs	1
BERT-based translation models (mBERT, XLM-R, T5)	1

7.Rule-based Models and Extraction Pipelines

Model	Frequency
Rule-based extraction pipelines	1
Third-party expert system, AIML-based models	1
Reverse Maximum Matching (RMM) for word segmentation	1
Conditional Random Fields (CRF) for entity recognition	1
Structured problem-solving methodology	1

8. Ranking Algorithms

Model	Frequency
BM25	1
TF-IDF	1
PageRank	1
ExpertiseRank	1
HITS	1
BM25, TF-IDF, Cosine Similarity	1

9. OpenAI and InstructGPT

Model	Frequency
InstructGPT (text-davinci-003)	1
GPT-3 (text-davinci-002, text-davinci-003)	1
ChatGPT	1
GPT-4	1
GPT-3.5, LLM-Enhanced Retrieval	1
ReTA LLM, GPT 3.5, GPT 4	1
GPT-2	1
GPT-3, T5, BART	1
Large Language Models (LLMs)	1