# Question Answering System

# With NLP

# Outline

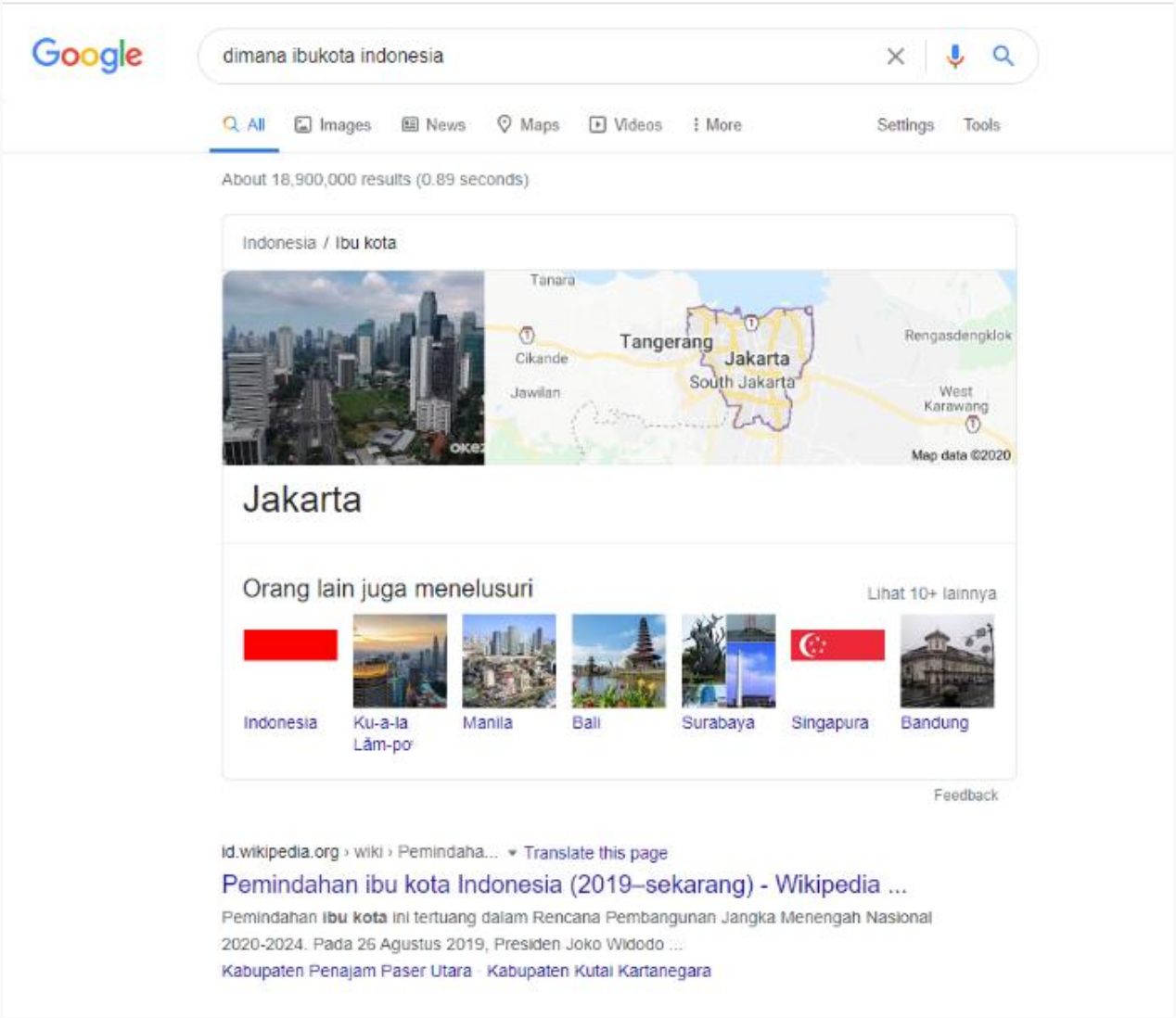
1. Introduction
2. Knowledge-based
3. Information Retrieval based
4. Machine Comprehension by Deep Learning
5. Conclusion

# Introduction

# What is Question Answering?

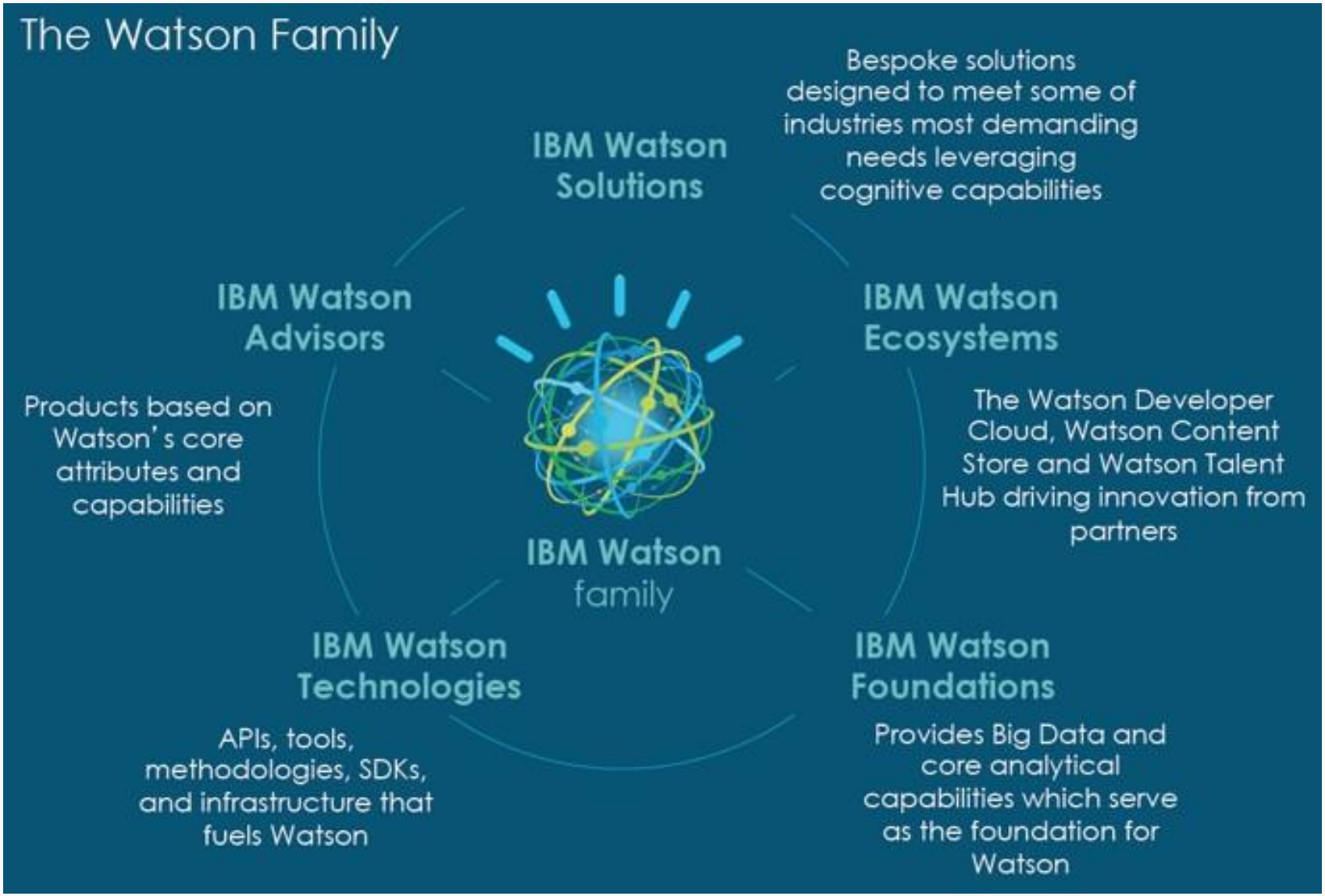
- **Answer particular questions**
- **Colloquial approach**
- **One of the oldest NLP problem**
- **A way to evaluate** Reading Comprehension
  - **"Since questions can be devised to query** any aspect **of text comprehension, the ability to answer questions is** the strongest possible demonstration of understanding**." Wendy Lehnert, 1977**

# Example - Google



# Example - IBM Waston

**Won Jeopardy on February 16, 2011!**



# Example - Siri

# Look Good?

# Type

- Single **vs Multiple**
- Simple **vs Complex**
- Text **vs Visual**
- Open-domain **vs Closed-domain**
- IR-based **vs Knowledge-based**

# Introduction

## Single vs Multiple

# Simple

- **A single document Q/A task involves questions associated with** one **particular document.**
- **In most cases, the assumption is that** the answer appears somewhere in the document **and probably once.**
- **Applications involve searching an individual resource, such as a book, encyclopedia, or manual.**
- **Reading comprehension tests are also a form of single document question answering.**
- **ex. SQuAD**

# Multiple

- **A multiple document Q/A task involves questions posed against** a collection of documents**.**
- **The answer may appear in the collection** multiple times **or** may not appear at all**!**
- **Applications include WWW search engines, and searching text repositories such as news archives, medical literature, or scientific articles.**
- **ex. MS MARCO**

# Introduction

## Simple vs Complex

# Question Type

- Simple (factoid) questions (most commercial systems)
  - Who wrote the Declaration of Independence?
  - What is the average age of the onset of autism?
  - Where is Apple Computer based?
  - ex. SQuAD
- Complex (narrative) questions
  - What do scholars think about Jefferson's position on dealing with pirates?
  - What is a Hajj?
  - In children with an acute febrile illness, what is the efficacy of single medication therapy with acetaminophen or ibuprofen in reducing fever?
  - ex. Narrative QA
- Complex (opinion) questions
  - Was the Gore/Bush election fair?

# Introduction

## Text vs Visual

# Text

**Input Document and Question.**

**Output Answer.**

**Example**

- **University of Washington**
- **AllenNlp**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

# Visual

**Input Picture or Video and Question.**

**Output Answer.**

- **What is in the image?**
- **Are there any humans?**
- **What sport is being played?**
- **Who has the ball?**
- **How many players are in the image?**
- **Who are the teams?**
- **Is it raining?**

**Example**

- **http://vqa.cloudcv.org**

# Introduction

## IR-based vs Knowledge-based

# Information Retrieval based

- **Information Retrieval:**
  **QA can be viewed as short passage retrieval.**
- **Information Extraction:**
  **QA can be viewed as open-domain information extraction.**

# Knowledge-based

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases
- Examples: Siri

# Hybrid based

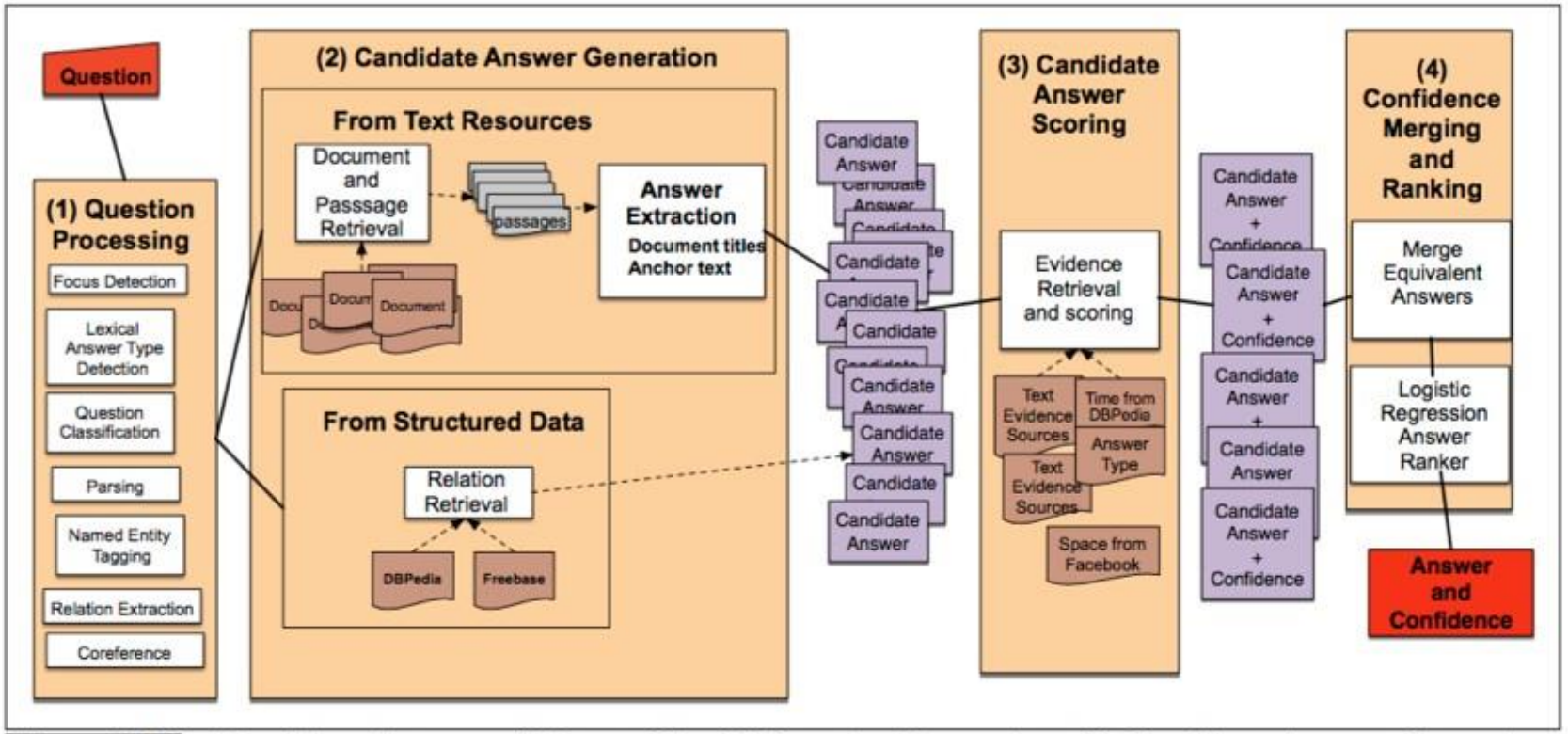



**Figure 28.9** The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

# Knowledge-based

# Knowledge-based

- **Semantic Pair**

| Question | Logical form |
|---|---|
| When was Ada Lovelace born? | `birth-year (Ada Lovelace, ?x)` |
| What states border Texas? | $\lambda$ x.state(x) $\wedge$ borders(x,texas) |
| What is the largest state | argmax($\lambda x$.state($x$), $\lambda x$.size($x$)) |
| How many people survived the sinking of the Titanic | `(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))` |

**Figure 23.9** Sample logical forms produced by a semantic parser for question answering. These range from simple relations like `birth-year`, or relations normalized to databases like Freebase, to full predicate calculus.

- **Data Base**

| subject | predict | object |
|---|---|---|
| **Ada Lovelace** | **birth-year** | **1815** |

# Knowledge-based (cont.)

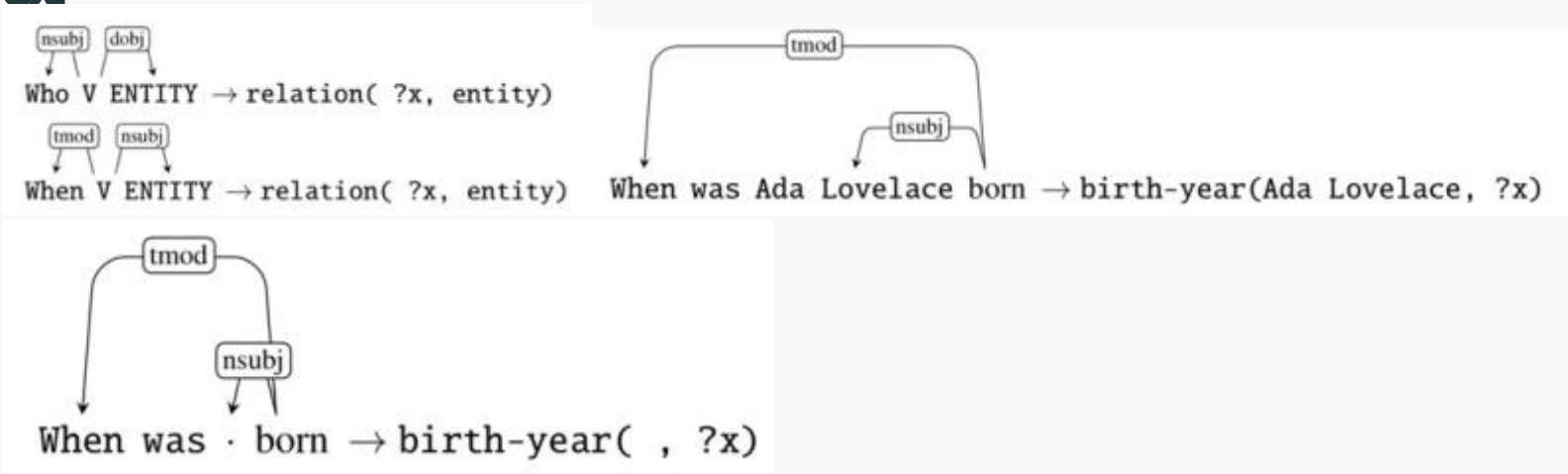
- **Rule-based Methods**

  **For example, to extract the birth-year relation, we could write patterns that search for the** question word **When, a main verb like born and that extract the named entity argument of the verb.**

- **Supervised Methods**

  **Generally these systems bootstrap by having a small set of rules for building this mapping and an initial lexicon as well.**

  **ex.**



# Knowledge-based (cont.)

- **Semi-Supervised Methods Because it is difficult to create training sets with questions labeled with their meaning representation, supervised datasets can't cover the wide variety of forms that even simple factoid questions can take.**

**ex. phrase align relation**

| | | |
|---|---|---|
| capital of | capital city of | become capital of |
| capitol of | national capital of | official capital of |
| political capital of | administrative capital of | beautiful capital of |
| capitol city of | remain capital of | make capital of |
| political center of | bustling capital of | capital city in |
| cosmopolitan capital of | move its capital to | modern capital of |
| federal capital of | beautiful capital city of | administrative capital city of |

**Figure 23.10** Some phrases that align with the Freebase relation `country.capital` from Berant et al. (2013).

**ex. Machine Translation**

| **Q: What are the green blobs in plant cells?** |
|---|
| *Lemmatized synonyms from PARALEX:* |
| what be the green blob in plant cell? |
| what be green part in plant cell? |
| what be the green part of a plant cell? |
| what be the green substance in plant cell? |
| what be the part of plant cell that give it green color? |
| what cell part do plant have that enable the plant to be give a green color? |
| what part of the plant cell turn it green? |
| part of the plant cell where the cell get it green color? |
| the green part in a plant be call? |
| the part of the plant cell that make the plant green be call? |

# Information Retrieval based

# Information Retrieval based

1. **Question Processing**
2. **Passage Retrieval**
3. **Answer Processing**

# Question Processing

- **Answer type: the kind of entity the answer consists of (person, location, time, etc.)**
- **Query: the keywords that should be used for the IR system to use in searching for documents**
- **Focus: the string of words in the question that are likely to be replaced by the answer in any answer string found**
- **ex. Which US state capital has the largest population?**
  query: **"US state capital has the largest population"**
  answer type: **city**
  focus: **state capital**

- **Query formulation is the task of creating a query-a list of tokens- to send to an information retrieval system to retrieve documents that might contain answer strings.**
- **"when was the laser invented?" might be reformulated as "the laser was invented"**
- **"where is the Valley of the Kings?" as "the Valley of the Kings is located in"**
- **hand-written rule**

  **ex.**

  wh-word **did A** verb **B** → **A** verb+ed **B**

  **Where is A → A is located in**

# Answer Type

- **Some systems make use of question classification, the task of finding the answer type, the named-entity categorizing the answer.**

**"Who founded Virgin Airlines?" expects an answer of type PERSON.**
**"What Canadian city has the largest population?" expects an answer of type CITY.**

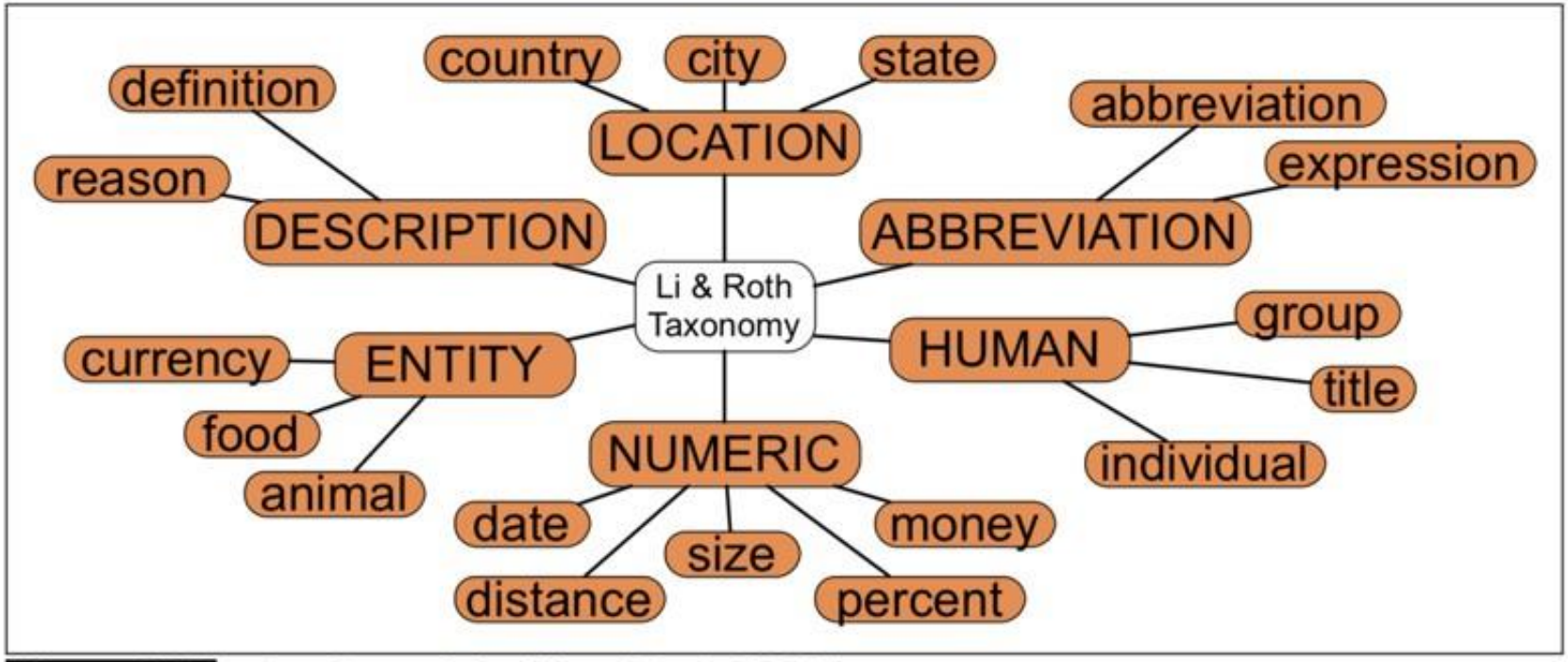



**Figure 23.3** A subset of the Li and Roth (2005) answer types.

# Passage Retrieval

- The number of named entities of the right type in the passage
- The number of question keywords in the passage
- The longest exact sequence of question keywords that occurs in the passage
- The rank of the document from which the passage was extracted
- The proximity of the keywords from the original query to each other
- The number of n-grams that overlap between the passage and the question

# Answer Extraction

- **Feature-based Answer Extraction**
- **N-gram tiling answer Extraction**
- **Neural Answer Extraction**

**"Who is the prime minister of India?"**

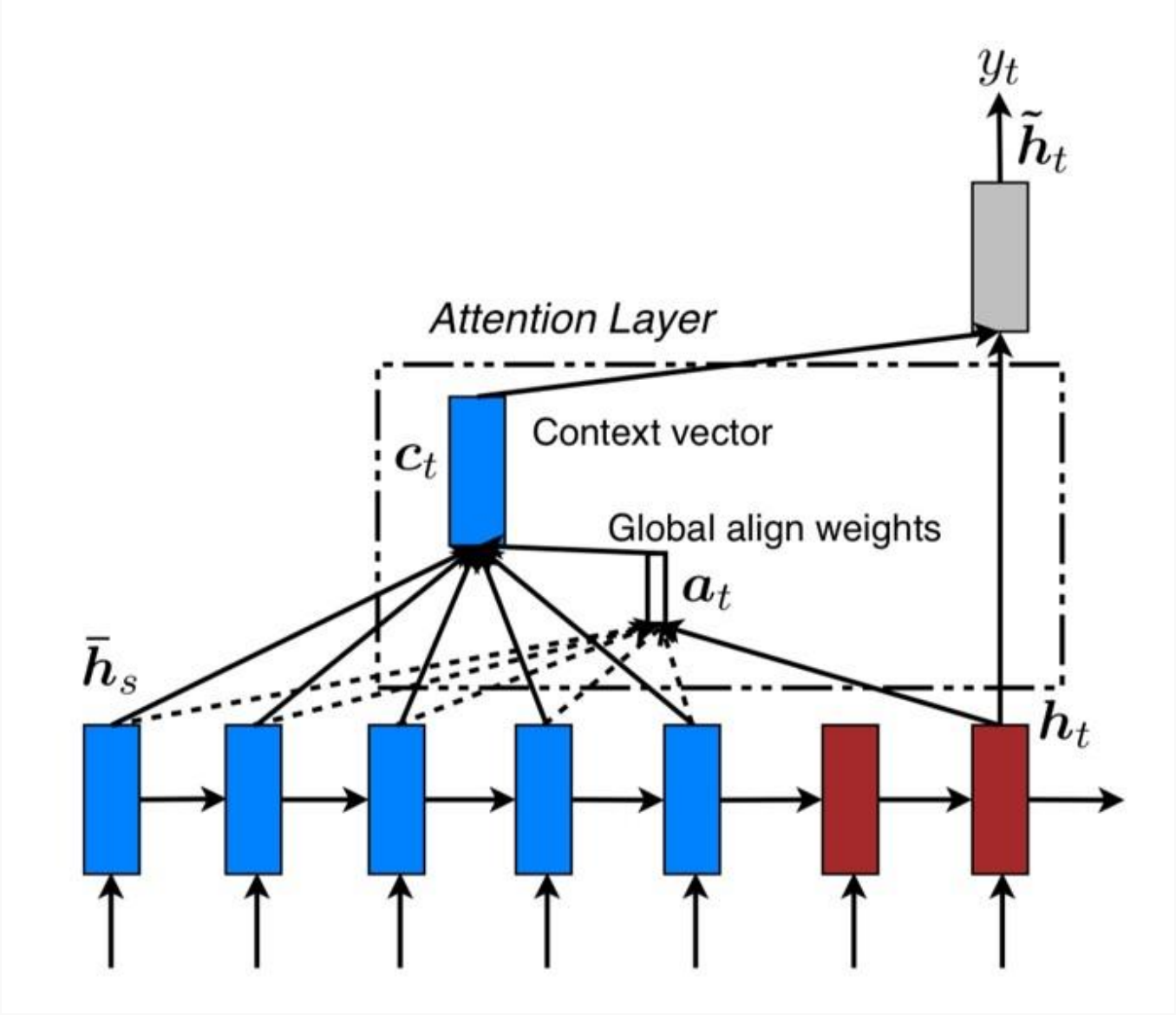
*Manmoham Singh, Prime Minsiter of India, had told left leaders that the deal would not be renegotiated.*

**"How tall is Mt. Everest?"**

*The official height of Mount Everest is 29029 feet*

# Machine Comprehension by Deep Learning

# Recap: Attention



# Machine Comprehension by Deep Learning

# End-To-End Memory Networks

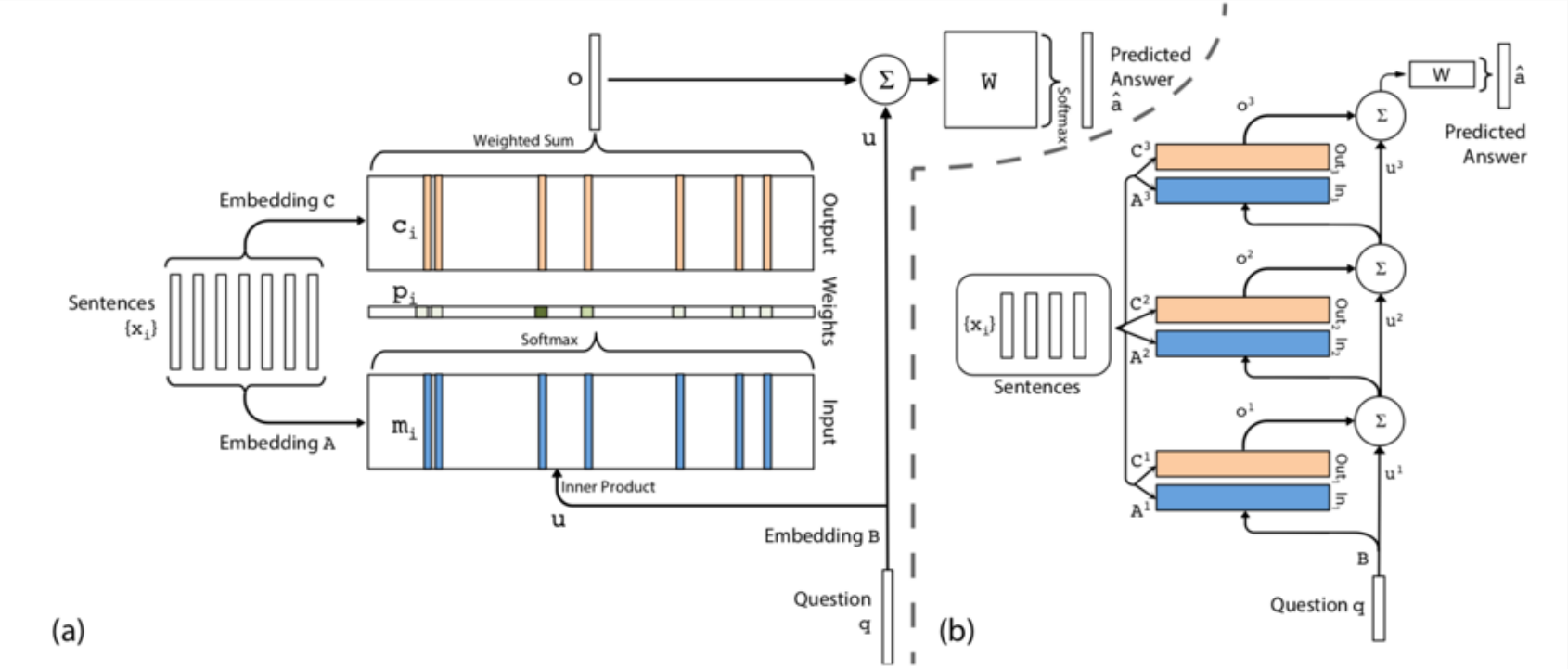
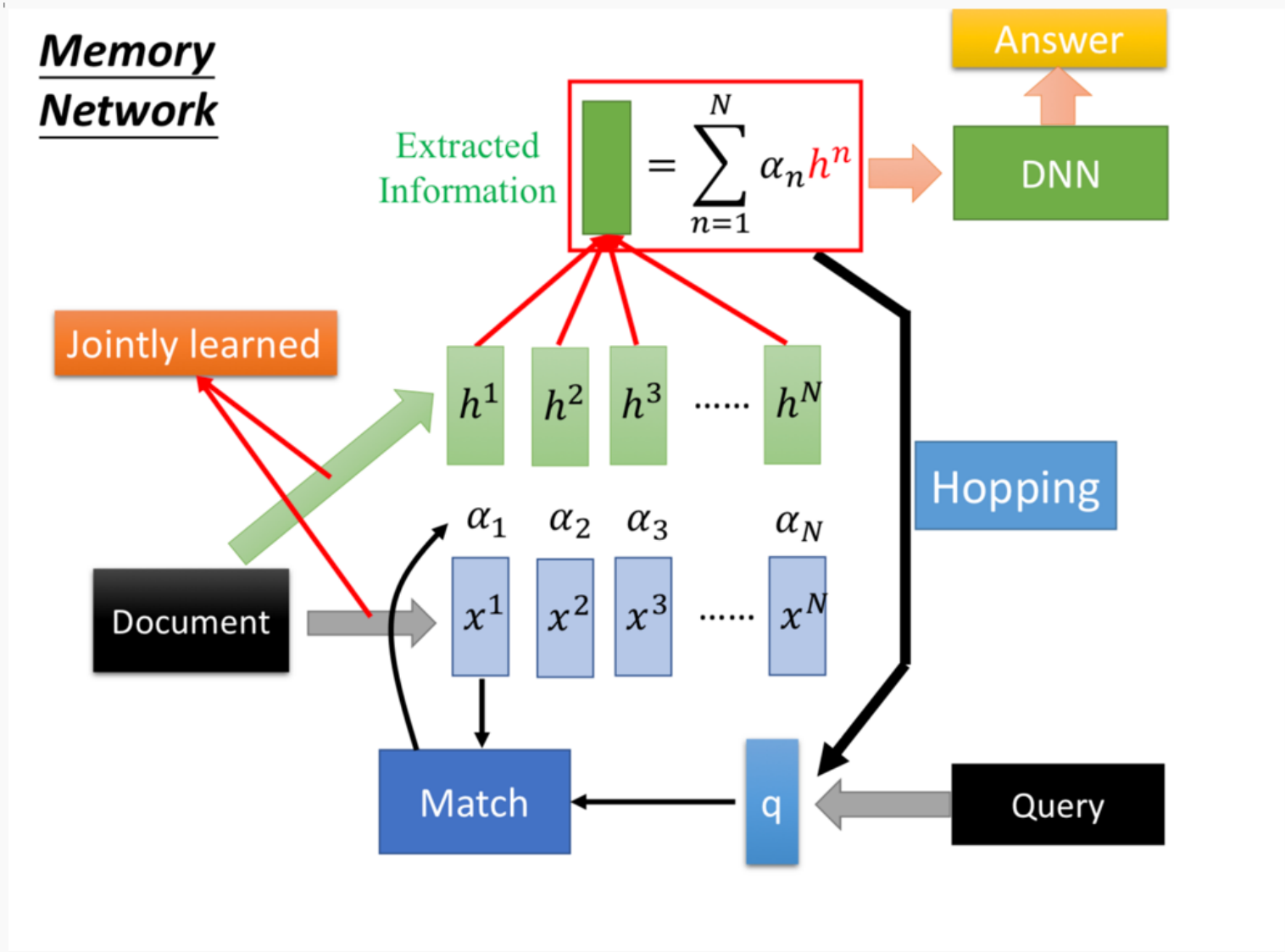


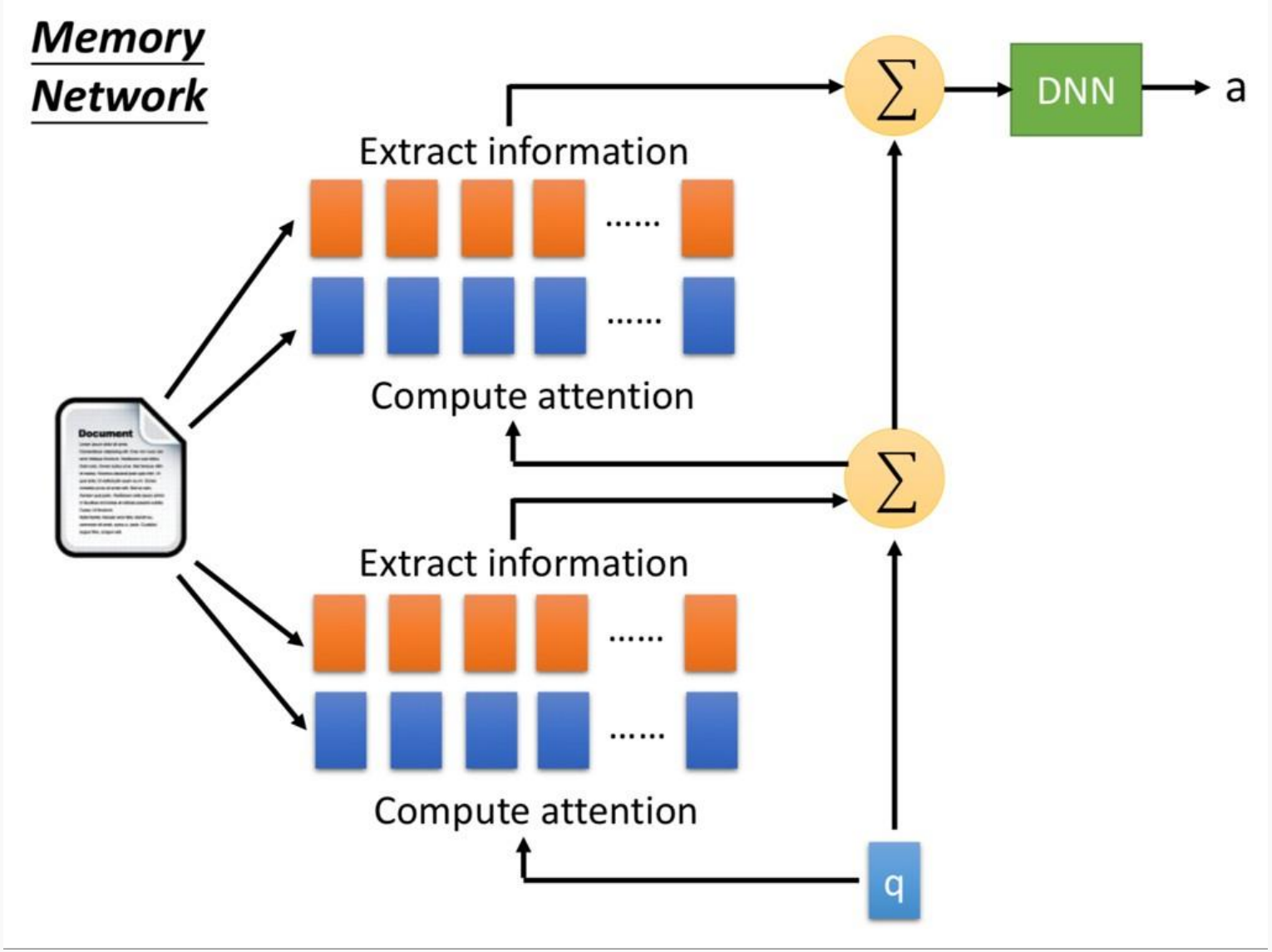

Figure 1: (a): A single layer version of our model. (b): A three layer version of our model. In practice, we can constrain several of the embedding matrices to be the same (see Section 2.2).

# End-To-End Memory Networks (cont.)




感謝李宏毅教授提供

# End-To-End Memory Networks (cont.)




感謝李宏毅教授提供

# End-To-End Memory Networks (cont.)

| Story (1: 1 supporting fact) | Support | Hop 1 | Hop 2 | Hop 3 |
|---|---|---|---|---|
| Daniel went to the bathroom. | | 0.00 | 0.00 | 0.03 |
| Mary travelled to the hallway. | | 0.00 | 0.00 | 0.00 |
| John went to the bedroom. | | 0.37 | 0.02 | 0.00 |
| John travelled to the bathroom. | yes | 0.60 | 0.98 | 0.96 |
| Mary went to the office. | | 0.01 | 0.00 | 0.00 |
| **Where is John? Answer: bathroom Prediction: bathroom** | | | | |

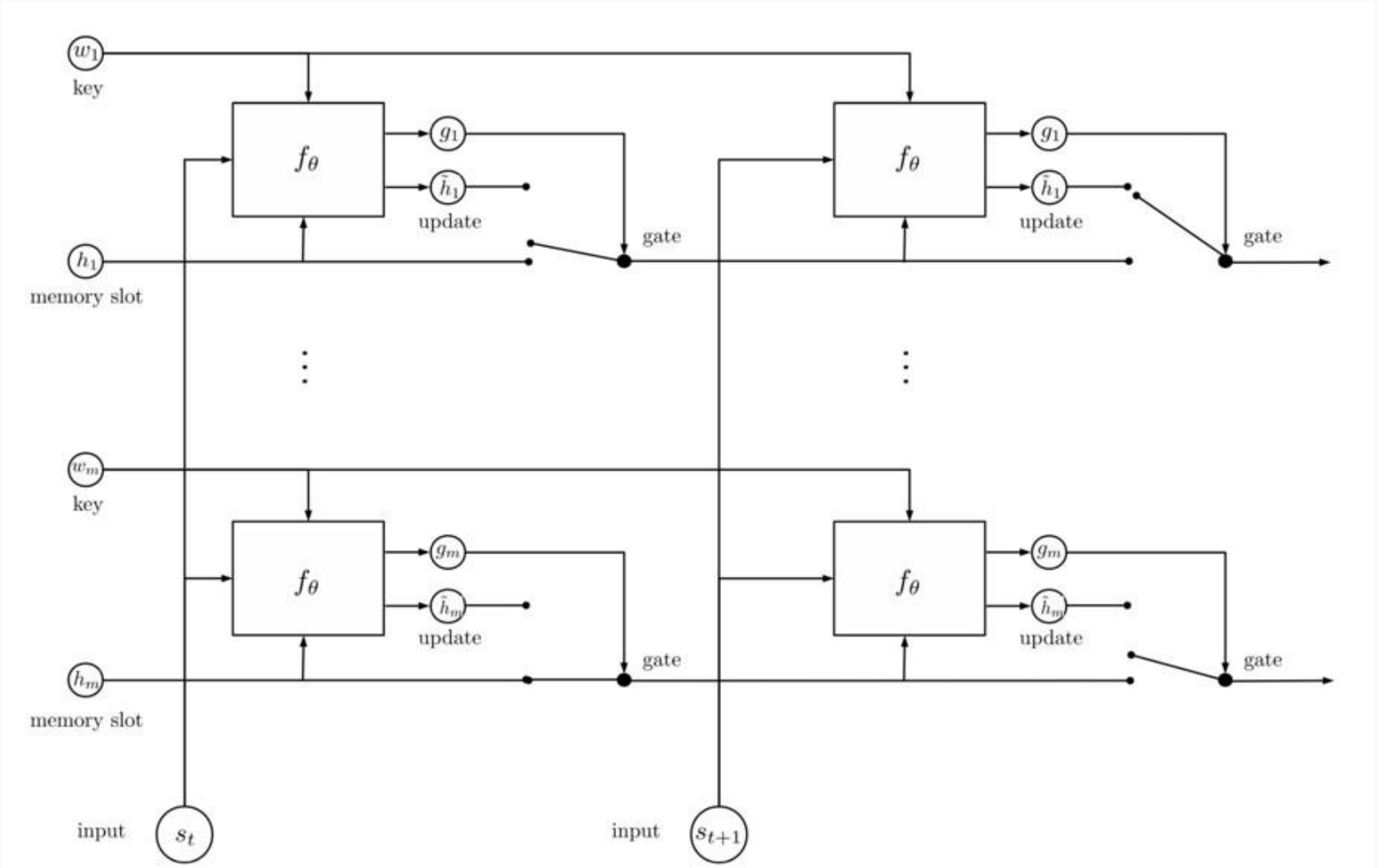
| Story (2: 2 supporting facts) | Support | Hop 1 | Hop 2 | Hop 3 |
|---|---|---|---|---|
| John dropped the milk. | | 0.06 | 0.00 | 0.00 |
| John took the milk there. | yes | 0.88 | 1.00 | 0.00 |
| Sandra went back to the bathroom. | | 0.00 | 0.00 | 0.00 |
| John moved to the hallway. | yes | 0.00 | 0.00 | 1.00 |
| Mary went back to the bedroom. | | 0.00 | 0.00 | 0.00 |
| **Where is the milk? Answer: hallway Prediction: hallway** | | | | |

| Story (16: basic induction) | Support | Hop 1 | Hop 2 | Hop 3 |
|---|---|---|---|---|
| Brian is a frog. | yes | 0.00 | 0.98 | 0.00 |
| Lily is gray. | | 0.07 | 0.00 | 0.00 |
| Brian is yellow. | yes | 0.07 | 0.00 | 1.00 |
| Julius is green. | | 0.06 | 0.00 | 0.00 |
| Greg is a frog. | yes | 0.76 | 0.02 | 0.00 |
| **What color is Greg? Answer: yellow Prediction: yellow** | | | | |

| Story (18: size reasoning) | Support | Hop 1 | Hop 2 | Hop 3 |
|---|---|---|---|---|
| The suitcase is bigger than the chest. | yes | 0.00 | 0.88 | 0.00 |
| The box is bigger than the chocolate. | | 0.04 | 0.05 | 0.10 |
| The chest is bigger than the chocolate. | yes | 0.17 | 0.07 | 0.90 |
| The chest fits inside the container. | | 0.00 | 0.00 | 0.00 |
| The chest fits inside the box. | | 0.00 | 0.00 | 0.00 |
| **Does the suitcase fit in the chocolate? Answer: no Prediction: no** | | | | |

| | Baseline | | | MemN2N | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | Strongly Supervised MemNN [22] | LSTM [22] | MemNN WSH | BoW | PE | PE LS | PE LS RN | 1 hop PE LS joint | 2 hops PE LS joint | 3 hops PE LS joint | PE LS RN joint | PE LS LW joint |
| 1: 1 supporting fact | 0.0 | 50.0 | 0.1 | 0.6 | 0.1 | 0.2 | 0.0 | 0.8 | 0.0 | 0.1 | 0.0 | 0.1 |
| 2: 2 supporting facts | 0.0 | 80.0 | 42.8 | 17.6 | 21.6 | 12.8 | 8.3 | 62.0 | 15.6 | 14.0 | 11.4 | 18.8 |
| 3: 3 supporting facts | 0.0 | 80.0 | 76.4 | 71.0 | 64.2 | 58.8 | 40.3 | 76.9 | 31.6 | 33.1 | 21.9 | 31.7 |
| 4: 2 argument relations | 0.0 | 39.0 | 40.3 | 32.0 | 3.8 | 11.6 | 2.8 | 22.8 | 2.2 | 5.7 | 13.4 | 17.5 |
| 5: 3 argument relations | 2.0 | 30.0 | 16.3 | 18.3 | 14.1 | 15.7 | 13.1 | 11.0 | 13.4 | 14.8 | 14.4 | 12.9 |
| 6: yes/no questions | 0.0 | 52.0 | 51.0 | 8.7 | 7.9 | 8.7 | 7.6 | 7.2 | 2.3 | 3.3 | 2.8 | 2.0 |
| 7: counting | 15.0 | 51.0 | 36.1 | 23.5 | 21.6 | 20.3 | 17.3 | 15.9 | 25.4 | 17.9 | 18.3 | 10.1 |
| 8: lists/sets | 9.0 | 55.0 | 37.8 | 11.4 | 12.6 | 12.7 | 10.0 | 13.2 | 11.7 | 10.1 | 9.3 | 6.1 |
| 9: simple negation | 0.0 | 36.0 | 35.9 | 21.1 | 23.3 | 17.0 | 13.2 | 5.1 | 2.0 | 3.1 | 1.9 | 1.5 |
| 10: indefinite knowledge | 2.0 | 56.0 | 68.7 | 22.8 | 17.4 | 18.6 | 15.1 | 10.6 | 5.0 | 6.6 | 6.5 | 2.6 |
| 11: basic coreference | 0.0 | 38.0 | 30.0 | 4.1 | 4.3 | 0.0 | 0.9 | 8.4 | 1.2 | 0.9 | 0.3 | 3.3 |
| 12: conjunction | 0.0 | 26.0 | 10.1 | 0.3 | 0.3 | 0.1 | 0.2 | 0.4 | 0.0 | 0.3 | 0.1 | 0.0 |
| 13: compound coreference | 0.0 | 6.0 | 19.7 | 10.5 | 9.9 | 0.3 | 0.4 | 6.3 | 0.2 | 1.4 | 0.2 | 0.5 |
| 14: time reasoning | 1.0 | 73.0 | 18.3 | 1.3 | 1.8 | 2.0 | 1.7 | 36.9 | 8.1 | 8.2 | 6.9 | 2.0 |
| 15: basic deduction | 0.0 | 79.0 | 64.8 | 24.3 | 0.0 | 0.0 | 0.0 | 46.4 | 0.5 | 0.0 | 0.0 | 1.8 |
| 16: basic induction | 0.0 | 77.0 | 50.5 | 52.0 | 52.1 | 1.6 | 1.3 | 47.4 | 51.3 | 3.5 | 2.7 | 51.0 |
| 17: positional reasoning | 35.0 | 49.0 | 50.9 | 45.4 | 50.1 | 49.0 | 51.0 | 44.4 | 41.2 | 44.5 | 40.4 | 42.6 |
| 18: size reasoning | 5.0 | 48.0 | 51.3 | 48.1 | 13.6 | 10.1 | 11.1 | 9.6 | 10.3 | 9.2 | 9.4 | 9.2 |
| 19: path finding | 64.0 | 92.0 | 100.0 | 89.7 | 87.4 | 85.6 | 82.8 | 90.7 | 89.9 | 90.2 | 88.0 | 90.6 |
| 20: agent's motivation | 0.0 | 9.0 | 3.6 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 |
| Mean error (%) | 6.7 | 51.3 | 40.2 | 25.1 | 20.3 | 16.3 | 13.9 | 25.8 | 15.6 | 13.3 | 12.4 | 15.2 |
| Failed tasks (err. > 5%) | 4 | 20 | 18 | 15 | 13 | 12 | 11 | 17 | 11 | 11 | 11 | 10 |
| On 10k training data | | | | | | | | | | | | |
| Mean error (%) | 3.2 | 36.4 | 39.2 | 15.4 | 9.4 | 7.2 | 6.6 | 24.5 | 10.9 | 7.9 | 7.5 | 11.0 |
| Failed tasks (err. > 5%) | 2 | 16 | 17 | 9 | 6 | 4 | 4 | 16 | 7 | 6 | 6 | 6 |

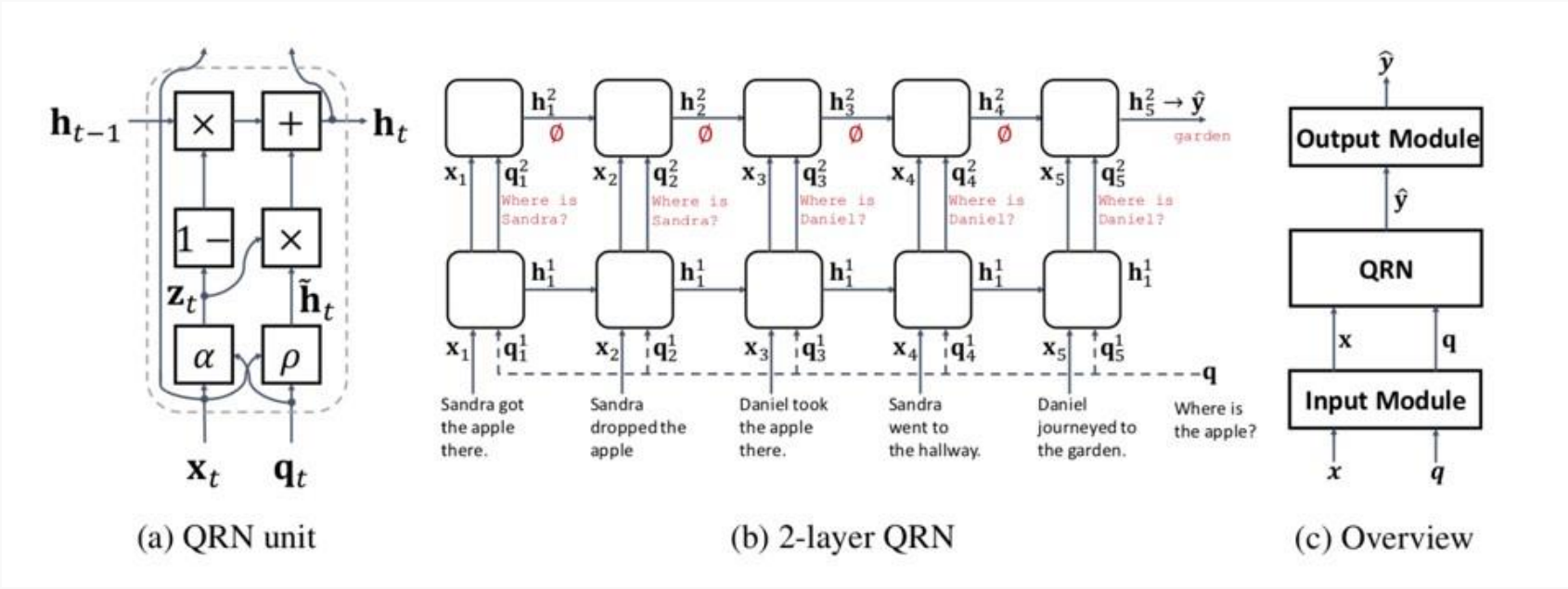
# Recurrent Entity Networks



# Recurrent Entity Networks (cont.)

| Task | NTM | D-NTM | MemN2N | DNC | DMN+ | EntNet |
|---|---|---|---|---|---|---|
| 1: 1 supporting fact | 31.5 | 4.4 | 0 | 0 | 0 | 0 |
| 2: 2 supporting facts | 54.5 | 27.5 | 0.3 | 0.4 | 0.3 | 0.1 |
| 3: 3 supporting facts | 43.9 | 71.3 | 2.1 | 1.8 | 1.1 | 4.1 |
| 4: 2 argument relations | 0 | 0 | 0 | 0 | 0 | 0 |
| 5: 3 argument relations | 0.8 | 1.7 | 0.8 | 0.8 | 0.5 | 0.3 |
| 6: yes/no questions | 17.1 | 1.5 | 0.1 | 0 | 0 | 0.2 |
| 7: counting | 17.8 | 6.0 | 2.0 | 0.6 | 2.4 | 0 |
| 8: lists/sets | 13.8 | 1.7 | 0.9 | 0.3 | 0.0 | 0.5 |
| 9: simple negation | 16.4 | 0.6 | 0.3 | 0.2 | 0.0 | 0.1 |
| 10: indefinite knowledge | 16.6 | 19.8 | 0 | 0.2 | 0 | 0.6 |
| 11: basic coreference | 15.2 | 0 | 0.0 | 0 | 0.0 | 0.3 |
| 12: conjunction | 8.9 | 6.2 | 0 | 0 | 0.2 | 0 |
| 13: compound coreference | 7.4 | 7.5 | 0 | 0 | 0 | 1.3 |
| 14: time reasoning | 24.2 | 17.5 | 0.2 | 0.4 | 0.2 | 0 |
| 15: basic deduction | 47.0 | 0 | 0 | 0 | 0 | 0 |
| 16: basic induction | 53.6 | 49.6 | 51.8 | 55.1 | 45.3 | 0.2 |
| 17: positional reasoning | 25.5 | 1.2 | 18.6 | 12.0 | 4.2 | 0.5 |
| 18: size reasoning | 2.2 | 0.2 | 5.3 | 0.8 | 2.1 | 0.3 |
| 19: path finding | 4.3 | 39.5 | 2.3 | 3.9 | 0.0 | 2.3 |
| 20: agent's motivation | 1.5 | 0 | 0 | 0 | 0 | 0 |
| Failed Tasks ($> 5\%$ error): | 16 | 9 | 3 | 2 | 1 | **0** |
| Mean Error: | 20.1 | 12.8 | 4.2 | 3.8 | 2.8 | **0.5** |

# Query-Reduction Networks




(a) QRN unit (b) 2-layer QRN (c) Overview

# Recurrent Entity Network (cont.)

| | Layer 1 | | | Layer 2 |
|---|---|---|---|---|
| Task 2: Two Supporting Facts | $z^1$ | $\overrightarrow{r}^1$ | $\overleftarrow{r}^1$ | $z^2$ |
| Sandra picked up the apple there. | 0.95 | 0.89 | 0.98 | 0.00 |
| Sandra dropped the apple. | 0.83 | 0.05 | 0.92 | 0.01 |
| Daniel grabbed the apple there. | 0.88 | 0.93 | 0.98 | 0.00 |
| Sandra travelled to the bathroom. | 0.01 | 0.18 | 0.63 | 0.02 |
| Daniel went to the hallway. | 0.01 | 0.24 | 0.62 | 0.83 |
| Where is the apple? | hallway | | | |

| | Layer 1 | | | Layer 2 |
|---|---|---|---|---|
| Task 15: Deduction | $z^1$ | $\overrightarrow{r}^1$ | $\overleftarrow{r}^1$ | $z^2$ |
| Mice are afraid of wolves. | 0.11 | 0.99 | 0.13 | 0.78 |
| Gertrude is a mouse. | 0.77 | 0.99 | 0.96 | 0.00 |
| Cats are afraid of sheep. | 0.01 | 0.99 | 0.07 | 0.03 |
| Winona is a mouse. | 0.14 | 0.85 | 0.77 | 0.05 |
| Sheep are afraid of wolves. | 0.02 | 0.98 | 0.27 | 0.05 |
| What is Gertrude afraid of? | wolf | | | |

| | Layer 1 | | | Layer 2 |
|---|---|---|---|---|
| Task 3: Displaying options | $z^1$ | $\overrightarrow{r}^1$ | $\overleftarrow{r}^1$ | $z^2$ |
| resto-paris-expen-frech-8stars? | 0.00 | 1.00 | 0.96 | 0.91 |
| Do you have something else? | 0.41 | 0.99 | 0.00 | 0.00 |
| Sure let me find another option. | 1.00 | 0.00 | 0.00 | 0.12 |
| resto-paris-expen-frech-5stars? | 0.00 | 1.00 | 0.96 | 0.91 |
| No this does not work for me. | 0.00 | 0.00 | 0.14 | 0.00 |
| Sure let me find an other option. | 1.00 | 0.00 | 0.00 | 0.12 |
| What do you think of this? resto-paris-expen-french-4stars | | | | |

| | Layer 1 | | | Layer 2 |
|---|---|---|---|---|
| Task 6: DSTC2 dialog | $z^1$ | $\overrightarrow{r}^1$ | $\overleftarrow{r}^1$ | $z^2$ |
| Spanish food. | 0.84 | 0.07 | 0.00 | 0.82 |
| You are lookng for a spanish restaurant right? | 0.98 | 0.02 | 0.49 | 0.75 |
| Yes. | 0.01 | 1.00 | 0.33 | 0.13 |
| What part of town do you have in mind? | 0.20 | 0.73 | 0.41 | 0.11 |
| I don't care. | 0.00 | 1.00 | 0.02 | 0.00 |
| What price range would you like? | 0.72 | 0.46 | 0.52 | 0.72 |
| I don't care. API CALL spanish R-location R-price | | | | |

| Task | 1k | | | | | | | | | | 10k | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Previous works | | | | QRN | | | | | | Previous works | | | QRN | | | |
| | LSTM | N2N | DMN+ | GMemN2N | 1r | 2 | 2r | 3r | 6r | 6r200* | N2N | DMN+ | GMemN2N | 2r | 2rv | 3r | 6r200 |
| 1: Single supporting fact | 50.0 | 0.1 | 1.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 13.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2: Two supporting facts | 80.0 | 18.8 | 72.3 | 8.1 | 65.7 | 1.2 | 0.7 | 0.5 | 1.5 | 15.3 | 0.3 | 0.3 | 0.0 | 0.4 | 0.8 | 0.4 | 0.0 |
| 3: Three supporting facts | 80.0 | 31.7 | 73.3 | 38.7 | 68.2 | 17.5 | 5.7 | 1.2 | 15.3 | 13.8 | 2.1 | 1.1 | 4.5 | 0.4 | 1.4 | 0.0 | 0.0 |
| 4: Two arg relations | 39.0 | 17.5 | 26.9 | 0.4 | 0.0 | 0.0 | 0.0 | 0.7 | 9.0 | 13.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5: Three arg relations | 30.0 | 12.9 | 25.6 | 1.0 | 1.0 | 1.1 | 1.1 | 1.2 | 1.3 | 12.5 | 0.8 | 0.5 | 0.2 | 0.5 | 0.2 | 0.3 | 0.0 |
| 6: Yes/no questions | 52.0 | 2.0 | 28.5 | 8.4 | 0.1 | 0.0 | 0.9 | 1.2 | 50.6 | 15.5 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7: Counting | 51.0 | 10.1 | 21.9 | 17.8 | 10.9 | 11.1 | 9.6 | 9.4 | 13.1 | 15.3 | 2.0 | 2.4 | 1.8 | 1.0 | 0.7 | 0.7 | 0.0 |
| 8: Lists/sets | 55.0 | 6.1 | 21.9 | 12.5 | 6.8 | 5.7 | 5.6 | 3.7 | 7.8 | 15.1 | 0.9 | 0.0 | 0.3 | 1.4 | 0.6 | 0.8 | 0.4 |
| 9 : Simple negation | 36.0 | 1.5 | 42.9 | 10.7 | 0.0 | 0.6 | 0.0 | 0.0 | 32.7 | 13.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 10: Indefinite knowledge | 56.0 | 2.6 | 23.1 | 16.5 | 0.8 | 0.6 | 0.0 | 0.0 | 3.5 | 12.9 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| 11: Basic coreference | 38.0 | 3.3 | 4.3 | 0.0 | 11.3 | 0.5 | 0.0 | 0.0 | 0.9 | 14.7 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 12: Conjunction | 26.0 | 0.0 | 3.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 15.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 13: Compound coreference | 6.0 | 0.5 | 7.8 | 0.0 | 5.3 | 5.5 | 0.0 | 0.3 | 8.9 | 13.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 14: Time reasoning | 73.0 | 2.0 | 61.9 | 1.2 | 20.2 | 1.3 | 0.8 | 3.8 | 18.2 | 14.5 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.1 |
| 15: Basic deduction | 79.0 | 1.8 | 47.6 | 0.0 | 39.4 | 0.0 | 0.0 | 0.0 | 0.1 | 14.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 16: Basic induction | 77.0 | 51.0 | 54.4 | 0.1 | 50.6 | 54.8 | 53.0 | 53.4 | 53.5 | 15.5 | 51.8 | 45.3 | 0.0 | 49.4 | 50.4 | 49.1 | 0.0 |
| 17: Positional reasoning | 49.0 | 42.6 | 44.1 | 41.7 | 40.6 | 36.5 | 34.4 | 51.8 | 52.0 | 13.0 | 18.6 | 4.2 | 27.8 | 0.9 | 0.0 | 5.8 | 4.1 |
| 18: Size reasoning | 48.0 | 9.2 | 9.1 | 9.2 | 8.2 | 8.6 | 7.9 | 8.8 | 47.5 | 14.9 | 5.3 | 2.1 | 8.5 | 1.6 | 8.4 | 1.8 | 0.7 |
| 19: Path finding | 92.0 | 90.6 | 90.8 | 88.5 | 88.8 | 89.8 | 78.7 | 90.7 | 88.6 | 13.6 | 2.3 | 0.0 | 31.0 | 36.1 | 1.0 | 27.9 | 0.1 |
| 20: Agents motivations | 9.0 | 0.2 | 2.2 | 0.0 | 0.0 | 0.0 | 0.2 | 0.3 | 5.5 | 14.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| # Failed | 20 | 10 | 16 | 10 | 12 | 8 | 7 | **5** | 13 | 20 | 3 | 1 | 3 | 2 | 2 | 3 | **0** |
| Average error rates (%) | 51.3 | 15.2 | 33.2 | 12.7 | 20.1 | 11.7 | **9.9** | 11.3 | 20.5 | 14.2 | 4.2 | 2.8 | 3.7 | 4.6 | 3.2 | 4.3 | **0.3** |