Motivation:
- Effects of intermediate retrieval performance on downstream are not well-studied.
- Relationship between intermediate retrieved facts and the final output will give insights on model development.

Datasets: HotpotQA & FEVER

**Background & Motivation**

**Machine Reading at Scale (MRS)** (Chen et al. 2017)

![Query](image)

**Model**

**Intermediate Facts**
- Mostly Not Evaluated
- Evaluated

**Output**

**Overall Method**

![Diagram](image)

**Input**
- Retrieval
  - [CLS] Query | [SEP] Context | [SEP]
  - QA
    - [CLS] yes/no Query | [SEP] Context | [SEP]

**Objective**

\[ J_{\text{query}} = - \sum \log(p(y)) - \sum \log(1 - p(y)) \]

**Ablation**

<table>
<thead>
<tr>
<th>Method</th>
<th>P-Level Retrieval</th>
<th>S-Level Retrieval</th>
<th>Answer Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
</tr>
<tr>
<td>Whole Pip.</td>
<td>35.17</td>
<td>87.93</td>
<td>50.25</td>
</tr>
<tr>
<td>Pip. w/o p-level</td>
<td>6.02</td>
<td>89.53</td>
<td>11.19</td>
</tr>
<tr>
<td>Pip. w/o s-level</td>
<td>35.17</td>
<td>87.92</td>
<td>50.25</td>
</tr>
</tbody>
</table>

**Table 3:** Ablation table over the paragraph-level and sentence-level neural retrieval sub-modules on HotpotQA.

<table>
<thead>
<tr>
<th>Method</th>
<th>P-Level Retrieval</th>
<th>S-Level Retrieval</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Pip.</td>
<td>94.15</td>
<td>48.84</td>
<td>91.23</td>
</tr>
<tr>
<td>Pip. w/o p-level</td>
<td>94.69</td>
<td>18.11</td>
<td>92.03</td>
</tr>
<tr>
<td>Pip. w/o s-level</td>
<td>94.15</td>
<td>48.84</td>
<td>91.23</td>
</tr>
</tbody>
</table>

**Table 4:** Ablation over the paragraph-level and sentence-level neural retrieval sub-modules on FEVER.

**Analysis**

**Figure 1:** System Overview.

**Figure 2:** 200 test results.

**Figure 3:** The results of EM for supporting fact, answer span prediction, facilitating our direct date systems are asked to select evidential sentence.

**Table 5:** System performance on different answer types. ‘PN’ is Proper Noun.

**Acknowledgement:** This work was supported by awards from Venika, Google, Facebook, Salesforce, and Adobe.

**Reference:**

- Danqi Chen, Adam Flach, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer spoken question answering.