**Main Idea: Offsite Prompt Tuning + Differential Privacy**

![Diagram of prompt tuning and differential privacy](image)

- **Private data**
  - X
  - Y

- **Local Private Prompt Tuning**
  - Trusted Environment
  - Local model
  - DP-OPT

- **Cloud Inference**
  - Diff-Private Instruction
  - Query

**Privatizing Offsite Prompt Tuning**

- **Definition (differential privacy):** A randomized algorithm \( M \) is \((\varepsilon, \delta)\)-DP if for every adjacent dataset \( D, D' \) and every output set \( S \in \text{range}(M) \),
  \[
  Pr[M(D) \in S] \leq e^{\varepsilon}Pr[M(D') \in S] + \delta
  \]

**Private Prompt Generation**

- A natural idea is to leverage "sample-and-aggregate" paradigm.
- Challenge: the token space could be very large.
- Solution: first reduce the domain space with some privacy budget (Limited Domain Mechanism [3]).

**Privacy Leakage from the tuned prompts**

<table>
<thead>
<tr>
<th>Source</th>
<th>Vicuna-7B</th>
<th>Llama-2-70B</th>
<th>DaVinci-003</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>92.8(0.2)</td>
<td>93.3(1.8)</td>
<td>92.7(0.3)</td>
</tr>
<tr>
<td>Tree</td>
<td>59.6(0.7)</td>
<td>65.2(7.5)</td>
<td>53.7(7.9)</td>
</tr>
<tr>
<td>Mpa</td>
<td>75.8(6.2)</td>
<td>78.0(2.3)</td>
<td>81.4(1.6)</td>
</tr>
<tr>
<td>Disaster</td>
<td>61.7(3.2)</td>
<td>73.1(1.6)</td>
<td>77.0(1.9)</td>
</tr>
</tbody>
</table>

**References**


**Experiment Results**

Our method can significantly outperform the technique of local DP training (PromptDPSSGD) due to the transferability to larger LLMs. Our method is also comparable to the non-private algorithm (DNL-1).

**Below are some examples of generated private prompts.**

![Examples of generated private prompts](image)