A Appendices

A.1 Experimental Setup

The $\texttt{BERT}_{\text{base}}$ model and the $\texttt{RoBERTa}_{\text{BASE}}$ model use the same configuration. The two models both have 12 hidden transformer layers and 12 attention heads. The hidden size of the model is 768 and the intermediate size in the transformer layers is 3,072. The activation function in the transformer layers is `gelu`.

Pre-training The batch size of 32 sequences is used for pre-training. `Adam` with the learning rate of $5 \cdot 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, the `L2` weight decay of 0.01, the learning rate warm up over the first 10% steps, and the linear decay of the learning rate are used. A dropout probability of 0.1 is applied to all layers. The cross-entropy is used for the training loss of each task. For the masked language modeling tasks, the model is trained until the perplexity stops decreasing on the development set. For the other pre-training tasks, the model is trained until both the loss and the accuracy stop decreasing on the development set.

Fine-tuning For fine-tuning, the batch size and the optimization approach are the same as the pre-training. The dropout probability is always kept at 0.1. The training loss is the sum of the cross-entropy of two fine-tuning tasks as in §2.2.

A.2 Question Types Analysis

Tables in this section show the results with respect to the question types using all models (Section 3.2) in the order of performance.

Type	Dist.	EM	SM	UM
Where	18.16	68.3(±1.3)	78.8(±1.2)	89.2(±1.5)
When	13.57	63.8(±1.6)	75.2(±0.9)	86.0(±1.6)
What	18.48	54.1(±0.8)	72.5(±1.5)	84.0(±0.9)
Who	18.82	56.0(±1.3)	66.1(±1.3)	79.4(±1.2)
How	15.32	38.1(±0.7)	59.2(±1.6)	77.5(±0.7)
Why	15.65	32.0(±1.1)	56.0(±1.7)	68.5(±0.8)

Table 6: Results from `RoBERTa` by question types.

Type	Dist.	EM	SM	UM
Where	18.16	67.1(±1.2)	78.9(±0.6)	89.0(±1.1)
When	13.57	62.3(±0.7)	76.3(±1.3)	88.7(±0.9)
What	18.48	55.1(±0.8)	73.1(±0.8)	86.7(±0.8)
Who	18.82	56.2(±1.4)	64.0(±1.7)	77.1(±1.3)
How	15.32	41.2(±1.1)	61.2(±1.5)	79.8(±0.7)
Why	15.65	32.4(±0.7)	57.4(±0.8)	69.1(±1.4)

Table 7: Results from $\texttt{RoBERTa}_{\text{pre}}$ by question types.

Type	Dist.	EM	SM	UM
Where	18.16	66.1(±0.5)	79.9(±0.7)	89.8(±0.7)
When	13.57	63.3(±1.3)	76.4(±0.6)	88.9(±1.2)
What	18.48	56.4(±1.7)	74.0(±0.5)	87.7(±2.1)
Who	18.82	55.9(±0.8)	66.0(±1.7)	79.9(±1.1)
How	15.32	43.2(±2.3)	63.2(±2.5)	79.4(±0.7)
Why	15.65	33.3(±2.0)	57.3(±0.8)	69.8(±1.8)

Table 8: Results from $\texttt{RoBERTa}_{\text{our}}$ by question types.

Type	Dist.	EM	SM	UM
Where	18.16	57.3(±0.5)	70.2(±1.3)	79.4(±0.9)
When	13.57	56.1(±1.1)	69.7(±1.6)	78.6(±1.7)
What	18.48	45.0(±1.4)	64.4(±0.7)	77.0(±1.0)
Who	18.82	46.9(±1.1)	56.2(±1.4)	67.6(±1.4)
How	15.32	29.3(±0.8)	48.4(±1.2)	60.9(±0.7)
Why	15.65	23.4(±1.6)	46.1(±0.9)	56.4(±1.3)

Table 9: Results from `BERT` by question types.

Type	Dist.	EM	SM	UM
Where	18.16	62.8(±1.8)	72.3(±0.8)	82.1(±0.7)
When	13.57	60.7(±1.5)	70.7(±1.8)	80.4(±1.1)
What	18.48	43.2(±1.3)	64.3(±1.7)	75.6(±1.8)
Who	18.82	47.8(±1.1)	56.9(±1.9)	69.7(±0.7)
How	15.32	33.2(±1.3)	48.3(±0.6)	59.8(±1.1)
Why	15.65	22.9(±1.6)	46.6(±0.7)	54.9(±0.9)

Table 10: Results from $\texttt{BERT}_{\text{pre}}$ by question types.

Type	Dist.	EM	SM	UM
Where	18.16	63.3(±1.2)	72.9(±1.7)	77.0(±1.2)
When	13.57	48.4(±1.9)	66.5(±0.8)	79.5(±1.5)
What	18.48	52.1(±0.7)	69.2(±1.1)	81.3(±0.7)
Who	18.82	51.3(±1.1)	61.9(±0.9)	67.5(±0.9)
How	15.32	30.9(±0.9)	52.1(±0.7)	65.4(±1.1)
Why	15.65	29.2(±1.6)	53.2(±1.3)	65.7(±0.8)

Table 11: Results from $\texttt{BERT}_{\text{our}}$ by question types.

A.3 Error Examples

Each table in this section gives an error example from the excerpt. The gold answers are indicated by the solid underlines whereas the predicted answers are indicated by the wavy underlines.

Q	Why is Joey planning a big party?
J	Oh, we're having a big party tomorrow night. Later!
R	Whoa! Hey-hey, you planning on inviting us?
J	Nooo, later.
P	Hey!! Get your ass back here, Tribbiani!!
R	Hormones!
M	What Phoebe meant to say was umm, how come you're having a party and we're not invited?
J	Oh, it's Ross' bachelor party.
M	Sooo?

Table 12: An error example for the `why` question (Q). J: Joey, R: Rachel, P: Pheobe, M: Monica.

Q	Who opened the vent?
R	Ok, got the vent open.
P	Hi, I'm Ben. I'm hospital worker Ben. It's Ben... to the rescue!
R	Ben, you ready? All right, gimme your foot. Ok, on three, Ben. One, two, three. Ok, That's it, Ben.
-	(*Ross and Susan lift Phoebe up into the vent.*)
S	What do you see?
P	Well, Susan, I see what appears to be a dark vent. Wait. Yes, it is in fact a dark vent.
-	(*A janitor opens the closet door from the outside.*)

Table 13: An error example for the `who` question (Q). P: Pheobe, R: Ross, S: Susan.

Q	How does Joey try to convince the girl to hang out with him?
J	Oh yeah-yeah. And I got the duck totally trained. Watch this. Stare at the wall. Hardly move. Be white.
G	You are really good at that. So uh, I had fun tonight, you throw one hell of a party.
J	Oh thanks. Thanks. It was great meetin' ya. And listen if any of my friends gets married, or have a birthday, ...
G	Yeah, that would be great. So I guess umm, good night.
J	Oh unless you uh, you wanna hang around.
G	Yeah?
J	Yeah. I'll let you play with my duck.

Table 14: An error example for the `how` question (Q). J: Joey, G: The Girl.