A Appendix

A.1 Supplementary plot: OOD vs in-distribution on training dynamics information (Training and in-dis: RTE; OOD: WNLI)

Figure 3: Training cartography maps (training set: RTE). The number of heuristics related samples in RTE is small.

A.2 Supplementary plot: OOD vs in-distribution on syntactic characteristics (entailment)

Figure 4: Results for hypothesis 2.2. Training and in-distribution test samples are RTE, and OOD samples are WNLI.
Results presented are at the end of epoch 8 for MNLI training and the end of epoch 50 for RTE training. This is based on the epoch in which the training error has converged (around 0.02).

A.3 Supplementary plot: OOD vs in-distribution on training dynamics information (Training and in-dis: MNLI; OOD: RTE)

Figure 6: Training and evaluation cartography maps (train: MNLI, evaluation: RTE). The number of heuristics related samples in RTE is small.
A.4 Supplementary plot: OOD vs in-distribution on syntactic characteristics (non-entailment)

This section shows plots for correlation between confidence scores ($\hat{\mu}_i$) of non-entailment samples and $m_2$

(a) Train & in-distribution: MNLI, OOD: WNLI
(b) Train & in-distribution: RTE, OOD: WNLI
(c) Train & in-distribution: MNLI, OOD: RTE

Figure 7: Supplementary results for 3.2. Correlation between $\hat{\mu}_i$ of non-entailment samples and $m_2$

A.5 Supplementary material: Extra lexical overlap measure

We also added another measure to quantify tendency to adopt lexical overlap heuristic. We calculated $m_1 = \frac{|s_1 \cap s_2|}{|s_1|}$. Essentially, this measures how much percentage of words found in the premise ($s_1$) can also be found in the hypothesis ($s_2$).

(a) Correlation between $m_2$ and all samples $\hat{\mu}_i$
(b) Correlation between $m_2$ and entailment samples $\hat{\mu}_i$

Figure 8: Results for hypothesis 2.2. Training and in-distribution test samples are MNLI, and OOD samples are WNLI.

(a) Correlation between $m_2$ and all samples $\hat{\mu}_i$
(b) Correlation between $m_2$ and entailment samples $\hat{\mu}_i$

Figure 9: Results for hypothesis 2.2. Training and in-distribution test samples are RTE, and OOD samples are WNLI.
Figure 10: Results for hypothesis 2.2. Training and in-distribution test samples are MNLI, and OOD samples are RTE.