

A EXAMPLE OF ATTENTION MATRIX, BARCODES, AND FILTRATION

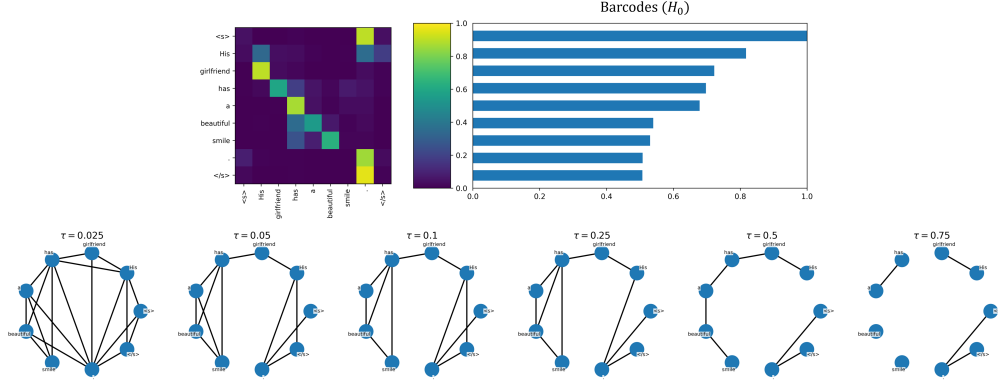


Figure 2: Attention matrix of RoBERTa-large-MNLI (layer 21, head 7; top left), H_0 barcode (top right), and the filtration of the attention graph indexed by the threshold τ (bottom). As τ increases, only edges with weights at least τ are retained, so the graph becomes progressively sparser. Input sentence: "His girlfriend has a beautiful smile." The bottom panel displays G_τ for $\tau \in \{0.025, 0.05, 0.1, 0.25, 0.5, 0.75\}$ in order.

B GRAPHICAL REPRESENTATION OF RTD-BARCODES

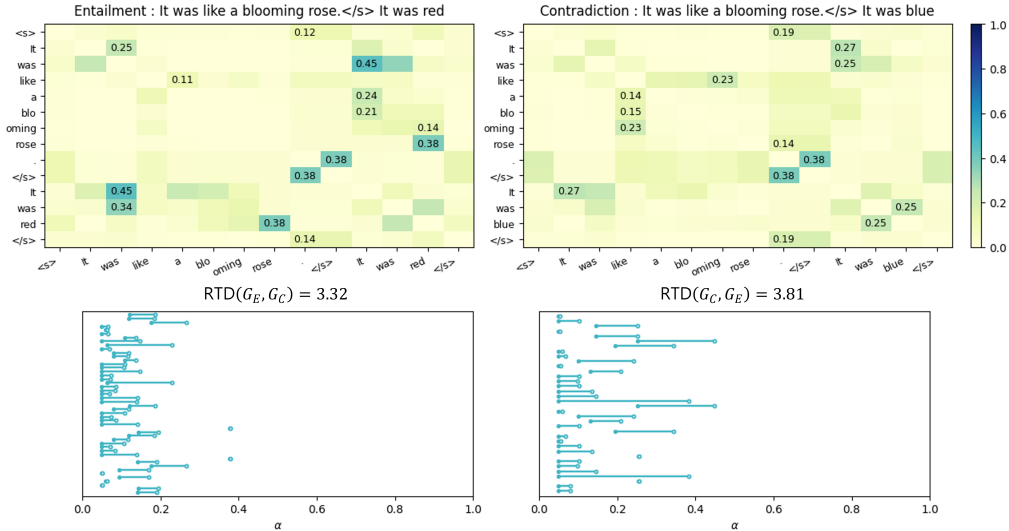


Figure 3: Example of RTD barcode construction. We use a sentence from the FLUTE Simile dataset. The top panels show self-attention matrices for Entailment (E) and Contradiction (C), with color indicating weight magnitude. The bottom panels display barcodes for $\text{RTD}(G_E, G_C)$ and $\text{RTD}(G_C, G_E)$. As the threshold α increases, for $\text{RTD}(A, B)$ a bar is born when an edge is present in B but absent in A , and it dies at the first α where that edge appears in both graphs. Each horizontal bar encodes the lifespan of the corresponding edge.

Figure 3 visualizes the RTD process. The scenario is “It was like a blooming rose.”, and the options are “It was red” (entailment) and “It was blue” (contradiction). The top panels show the self-attention matrices obtained by concatenating the scenario and option sequences from layer 19, head 0 of RoBERTa-large-MNLI. To compute RTD, we evaluate both $\text{RTD}(G_E, G_C)$ and $\text{RTD}(G_C, G_E)$. As the threshold α increases in $\text{RTD}(A, B)$, a bar is born when an edge is present in B but absent in A , and it dies at the first α where the edge appears in both graphs. The bottom panels display the corresponding barcodes. For visual clarity, only edges with attention weights at least 0.05 are shown. The formal definition applies over the entire range of α , and the RTD value is the sum of all bar lengths. The results are $\text{RTD}(G_E, G_C) = 3.32$ and $\text{RTD}(G_C, G_E) = 3.81$, so $\text{RTD}(G_E, G_C) < \text{RTD}(G_C, G_E)$ holds, indicating that the entailment side exhibits a more stable barcode structure and that the model selected an appropriate interpretation of the figurative expression.

C HEAD ENSEMBLE

We adapt the head ensemble strategy of Chakrabarty et al. (2022) to improve RTD in FLUTE. We decouple exploration from selection: candidate ensembles are scored on the training split, and for each ensemble size, the top candidate is validated on a held-out split to update the best-sofar model. Scoring uses soft voting with a fixed decision threshold $\tau = 0.5$: for each instance we average the scores of the rules included in the candidate. If this average exceeds τ , the ensemble issues the corresponding prediction. For each candidate, its score is the fraction of training instances correctly classified under this decision rule. We perform beam search with fixed width B , retaining the top B candidates at each step and accepting only expansions that do not decrease the score. By default, we set the maximum number of layer-head pairs per ensemble to $M = 64$ and the beam width to $B = 40$. The algorithm is shown in Algorithm 1

Algorithm 1 Head Ensemble

Require: $X, \tilde{X} \in \mathbb{R}^{N \times 2LH}$, L (layers), H (heads), B (beam width), M (max # of layer-head indices), $V \leftarrow LH$
 Defaults: $B = 40$, $M = 64$ (pairs are added, so steps = $\lfloor M/2 \rfloor$)

```

1: function SCORE( $S$ )
2:   return  $\frac{1}{N} \sum_{n=1}^N \left\{ \frac{1}{|S|} \sum_{v \in S} \tilde{X}_{n,v} > 0.5 \right\}$ 
3: end function
4:  $\mathcal{F} \leftarrow \{(\text{Score}(\{v\}), \{v\}) \mid v \in [0, V)\}$ ,  $S^* \leftarrow \emptyset$ ,  $best\_acc \leftarrow -\infty$ 
5: for  $t = 1$  to  $\lfloor M/2 \rfloor$  do ▷ add a pair per step
6:    $\mathcal{C} \leftarrow \emptyset$ 
7:   for all  $(s, S) \in \mathcal{F}$  do
8:     for all unordered pairs  $\{v_1, v_2\} \subset [0, V) \setminus S$ ,  $|v_1 - v_2| \neq LH$ :
9:        $S' \leftarrow S \cup \{v_1, v_2\}$ ,  $s' \leftarrow \text{SCORE}(S')$ 
10:      if  $s' \geq s$  then add  $(s', S')$  to  $\mathcal{C}$ 
11:   end for
12:    $\mathcal{F} \leftarrow$  top- $B$  elements of  $\mathcal{C}$  by score
13:   if  $\mathcal{F} \neq \emptyset$  then
14:      $(\hat{s}, \hat{S}) \leftarrow \arg \max_{(s, S) \in \mathcal{F}}$ 
15:      $acc \leftarrow \frac{100}{N} \sum_{n=1}^N \left\{ \frac{1}{|\hat{S}|} \sum_{v \in \hat{S}} X_{n,v} > 0.5 \right\}$ 
16:     if  $acc > best\_acc$  then  $best\_acc \leftarrow acc$ ,  $S^* \leftarrow \hat{S}$ 
17:     end if
18:   end if
19: end for
20: return  $S^*$ 

```

D SATS: TOPOLOGICAL AND ALGORITHMIC DETAILS

D.1 EQUIVALENCE OF DEATH AND WIDEST PATH

Claim : $d(o) = \text{wp}([S], o)$

Proof. For an option vertex o , define

$$D(o) := \{\tau \in I : o \text{ and } [S] \text{ are still disconnected in } K_\tau / \sim\}, \quad d(o) := \sup D(o).$$

Also, define the widest-path value.

$$\text{wp}(u, v) := \max_{p: u \rightsquigarrow v} \min_{e \in p} w_e, \quad \text{wp}([S], o) := \max_{s \in S} \text{wp}(s, o).$$

Let $d(o) \geq \text{wp}([S], o)$. For any $\tau > \text{wp}([S], o)$, every $s \in S$ and every path $p : s \rightsquigarrow o$ satisfy $\min_{e \in p} w_e < \tau$, so there is no path from $[S]$ to o in G_τ . Hence $\tau \in D(o)$, and since every $\tau > \text{wp}([S], o)$ lies in $D(o)$, we have $\sup D(o) \geq \text{wp}([S], o)$.

Let $d(o) \leq \text{wp}([S], o)$. For any $\tau < \text{wp}([S], o)$, by definition there exist $s^* \in S$ and a path $p^* : s^* \rightsquigarrow o$ with $\min_{e \in p^*} w_e \geq \tau$. Consequently, every edge of p^* lies in E_τ , so $[S]$ and o are connected in G_τ . Thus $\tau \notin D(o)$, and because no $\tau < \text{wp}([S], o)$ belongs to $D(o)$, we get $\sup D(o) \leq \text{wp}([S], o)$. Combining both directions yields $d(o) = \text{wp}([S], o)$. Therefore, death coincides with the graph widest path value. \square

D.2 LIPSCHITZ STABILITY AND LOWER BOUNDS FOR SATS

Claim : $u(o)$, $\text{SATS}(i)$ are Lipschitz.

Proof. Let U, V be symmetric matrices, then

$$U \preceq V \Rightarrow \Phi_K(U; \beta) \preceq \Phi_K(V; \beta),$$

$$\Phi_K(U; \beta) \preceq \Phi_{K+1}(U; \beta), \quad 0 < \beta_1 \leq \beta_2 < 1 \Rightarrow \Phi_K(U; \beta_1) \preceq \Phi_K(U; \beta_2).$$

Here, \preceq denotes entrywise order. If U increases entrywise, or if K or β increases, then Φ_K is entrywise monotonically increasing.

$$|u_U(o) - u_V(o)| \leq \|\Phi_K(U; \beta) - \Phi_K(V; \beta)\|_\infty,$$

$$|\text{SATS}_U(i) - \text{SATS}_V(i)| \leq \frac{1}{\varepsilon} \|\Phi_K(U; \beta) - \Phi_K(V; \beta)\|_\infty.$$

To justify the Lipschitz claim, we bound Φ_K . By the non-commutative telescoping series identity,

$$U^t - V^t = \sum_{j=0}^{t-1} U^{t-1-j} (U - V) V^j \quad (t \in \mathbb{N}),$$

hence

$$\Phi_K(U; \beta) - \Phi_K(V; \beta) = \sum_{t=1}^K \beta^{t-1} \sum_{j=0}^{t-1} U^{t-1-j} (U - V) V^j.$$

With the triangle inequality and submultiplicativity of the matrix norm,

$$\|\Phi_K(U; \beta) - \Phi_K(V; \beta)\| \leq \left(\sum_{t=1}^K t \beta^{t-1} M^{t-1} \right) \|U - V\|, \quad M = \max\{\|U\|, \|V\|\}.$$

In particular, on the norm-bounded domain $\{U : \|U\| \leq R\}$, $\Phi_K(\cdot; \beta)$ is $L_K(R, \beta)$ -Lipschitz with

$$L_K(R, \beta) = \sum_{t=1}^K t \beta^{t-1} R^{t-1}.$$

Therefore, $U \mapsto u(o)$ is $L_K(R, \beta)$ -Lipschitz and $U \mapsto \text{SATS}(i)$ is $\frac{1}{\varepsilon} L_K(R, \beta)$ -Lipschitz w.r.t. $\|\cdot\|_\infty$. \square

Claim : Φ_K has a lower bound.

Proof. For a lower link to widest paths, let $m = \text{wp}(s, o)$, and suppose a simple path achieving m has length t . Then

$$(\Phi_K(U; \beta))_{s,o} \geq \beta^{t-1} m^t \quad (t \leq K),$$

and, with the scenario contraction,

$$\max_{s \in S} (\Phi_K(U; \beta))_{s,o} \geq \max_{s \in S} \max_{1 \leq t \leq \min\{K, T-1\}} \beta^{t-1} \text{wp}(s, o)^t.$$

Therefore, whenever there is a strong widest path, Φ_K admits a corresponding lower bound and quantitatively reflects the strength of the widest path. \square

E DETAILS FOR LLMs

%% FLUTE Prompt (This sentence is not included) %%

You are a careful multiple choice grader for short story comprehension with figurative language.

You will receive:

- One instruction line.
- A short story and a question.
- Answer options labeled 1..N (N varies by item).

Rules:

- Use only the given text and commonsense.
- Prefer figurative/pragmatic meaning over literal.
- Do NOT invent options or rely on external facts.

Output:

- Output ONLY the chosen option number (1..N).
- One line, no spaces, no words, no punctuation.

{Hypothesis}

1. {Premise 1}
2. {Premise 2}

%% Pragmatics Prompt (This sentence is not included) %%

You are a careful multiple choice grader for short story comprehension with figurative language.

You will receive:

- One instruction line.
- A short story and a question.
- Answer options labeled 1..N (N varies by item).

Rules:

- Use only the given text and commonsense.
- Prefer figurative/pragmatic meaning over literal.
- Do NOT invent options or rely on external facts.

Output:

- Output ONLY the chosen option number (1..N).
- One line, no spaces, no words, no punctuation.

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple choice question. Read each story and choose the best answer to each question.

The answer options are 1, 2, 3, 4, or 5.

{Scenario}

1. {Option 1}
2. {Option 2}
3. {Option 3}
- ...
- N. {Option N}

To prevent diverse outputs and obtain only the index of the predicted option, we set temperature = 0.0, top_p = 1.0, max_new_tokens = 2, seed = 42. In addition, we suppress tokens such as “answer”, “correct”, and “option” through a negative lexicon so that the models output only the option index, and we restrict the first generated token to be a digit. After generation, any non-numeric output would have been excluded from evaluation. In practice, all models returned numeric responses.

In FLUTE, Hypothesis and Premise are dataset field names. Because each hypothesis has two premises, we remap the fields to align with the Pragmatics prompt format: Hypothesis → Scenario, Premise 1/2 → Option (1/2). When presenting premises as options, we do not preserve their original order. Instead, we shuffle them with a fixed random seed for reproducibility. The gold label is remapped to 1/2 according to the shuffled order and the model is asked to select the correct option (1/2) given the scenario. Only the field names and display order change and the underlying content remains unchanged.

F SC V2 IN FREEZE

Table 6: SC V2 results with TDA representation on Freeze.

Model	Overall		GEN		HYP		RQ	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
bert-base-uncased	76.5	77.3	66.0	66.5	73.4	73.4	72.0	72.4
RoBERTa-base	76.2	76.6	67.7	68.2	74.4	74.7	72.8	73.2
RoBERTa-large	78.6	78.9	68.6	68.8	75.9	76.2	74.4	74.6
DeBERTa-v3-base	78.5	78.9	69.1	70.0	76.3	76.4	74.6	75.1
DeBERTa-v3-large	80.7	80.9	74.3	74.0	74.0	74.1	76.3	76.3

Table 6 reports results of a TDA-based evaluation of five models under a frozen parameter setting on SC V2. All models achieve accuracy and F1 above the chance level of 50%. These findings suggest that even without fine-tuning, the attention structure of the models already encodes sarcasm-related signals, which can be exploited for classification via topological feature extraction alone. However, because the absolute performance is below that of fine-tuned models, we interpret these results primarily as evidence of signal presence rather than as a competitive alternative.

G RESULT OF RTD HEAD ENSEMBLE

RTD of Figure 5 shows the distribution of attention layer-head pairs selected by RoBERTa-large-MNLI and BART-large-MNLI when running the RTD head ensemble. We restricted the number of selectable layer-head matrices per exploration step to 40. Therefore, at most 40 selections are shown per task. Consistent with earlier analysis, RoBERTa primarily selected pairs from the mid-to-late layers, whereas BART distributed selections relatively evenly across layers. The number of selected pairs also varies by model and task. For example, on Metaphor, RoBERTa selects 3 pairs and BART selects 6 fewer than for the other tasks. BART selects more pairs on simile, whereas RoBERTa selects relatively more on Idiom. However, selection count is a byproduct of the ensemble search, rather than a proxy for accuracy, and it shows no consistent monotonic relationship to accuracy.

H ABLATION STUDY OF SATS

Table 7 reports the effect of multi-hop propagation in SATS on score computation. The same hyperparameters are held fixed across models and tasks. Only the aggregation (top- k , max, softmax) and the use of multi-hop are varied. For both models, switching the aggregation to max or softmax generally reduces accuracy or leaves it unchanged, indicating that top- k is the most consistent choice. Removing multi-hop (w/o multi-hop) typically lowers or maintains accuracy on DeBERTa-v3-large (notably, with top- k it drops on all tasks except Irony) and, on RoBERTa-large-MNLI with top- k , on all tasks except Pragmatics Metaphor. For RoBERTa, small gains are occasionally observed when multi-hop is removed under max or softmax. Overall, the multi-hop with top- k setting achieves the best performance. Regarding latency, disabling multi-hop always shortens runtime, because the computational complexity of Φ_K is $O(KT^3)$, linear in K , and turning off multi-hop removes K from the computation.

Table 7: Results Accuracy and Latency. Columns: Simile / Metaphor (FLUTE) / Idiom / Irony / Metaphor (Pragmatics). These results report the effect of varying the aggregation used to compute $u(o)$ with multi-hop propagation disabled. All other hyperparameters are held fixed across models and tasks.

Model	Simile	Metaphor	Idiom	Irony	Metaphor
DeBERTa-v3-large	81.6 / 4.67	91.9 / 2.86	97.6 / 3.62	68.0 / 3.95	65.0 / 1.88
w/o Multi-hop	77.6 / 1.69	90.3 / 1.61	96.0 / 2.03	68.0 / 0.99	55.0 / 0.76
Aggregation: Max	76.8 / 3.89	83.1 / 2.11	95.2 / 2.63	60.0 / 3.53	60.0 / 1.53
Max w/o Multi-hop	76.0 / 1.04	83.1 / 0.95	95.2 / 1.18	60.0 / 0.66	60.0 / 0.46
Aggregation: Softmax	80.8 / 4.48	84.7 / 2.63	96.0 / 3.24	68.0 / 3.81	60.0 / 1.77
Softmax w/o Multi-hop	77.6 / 1.60	83.9 / 1.47	96.0 / 1.79	64.0 / 0.95	55.0 / 0.70
RoBERTa-large-MNLI	79.2 / 6.30	92.7 / 5.19	97.6 / 4.64	80.0 / 3.31	55.0 / 3.32
w/o Multi-hop	76.0 / 1.83	85.5 / 1.53	97.6 / 1.99	72.0 / 1.02	55.0 / 0.74
Aggregation: Max	72.0 / 5.74	92.7 / 2.65	95.2 / 6.10	56.0 / 2.37	55.0 / 1.06
Max w/o Multi-hop	75.2 / 1.07	85.5 / 0.96	95.2 / 1.21	72.0 / 0.68	50.0 / 0.48
Aggregation: Softmax	78.4 / 6.35	91.1 / 3.18	95.2 / 6.74	64.0 / 2.66	55.0 / 1.31
Softmax w/o Multi-hop	76.8 / 1.64	82.3 / 1.64	96.0 / 1.82	68.0 / 0.97	55.0 / 0.72

In the latency experiments, for each comparison the hyperparameters were held at their default values: $K = 3$, $\beta = 0.6$, $\lambda = 0.6$, and top- $k = 3$. Across the sweeps, we vary K and top- k from 1 to 7, and β and λ from 0.3 to 0.9. Because the theoretical complexity is $O(KT^3)$, the remaining hyperparameters, excluding K , do not change the leading-order term of the latency. Figure 6 shows that the latency grows approximately linearly with K , consistent with the complexity analysis. In contrast, when multi-hop is disabled, K is unused in internal computations, so the latency remains nearly constant across settings.

Figure 4 shows accuracy as a function of the weighting coefficient $\lambda \in [0.3, 0.9]$ in the blend of multi-hop diffusion Φ_K and direct connections U . The left panel reports RoBERTa results, and the right panel DeBERTa results. All other settings ($K, \beta, \text{top-}k$) are fixed. The evaluation is performed in a deterministic zero-shot setup, and each point displays a 95% Wilson confidence interval. In general, accuracy decreases slightly as λ approaches either extreme. Apart from the Simile task with RoBERTa and the Metaphor (Pragmatics) task with DeBERTa, the variation is small, indicating robustness to λ . Consequently, we set the default to $\lambda \approx 0.6$. This indicates that SATS is stable with respect to λ and does not require extensive hyperparameter tuning.

I ADDITIONAL EXAMPLE OF SATS

In Figure 7 (B), Option 1 first links s_5 (rocker) to o_3 (insane). While rocker literally denotes a rocking chair, the idiom “be off one’s rocker” figuratively means “mentally unstable”. Therefore, it is semantically aligned with insane. By contrast, Option 2 contains sane and level-headed, which convey opposite meanings, but they do not connect to rocker. Instead, s_0 (You) and s_4 (your), which refer to the same entity, connect to o_0 (You) first. Consequently, Option 1 yields a lower SATS value.

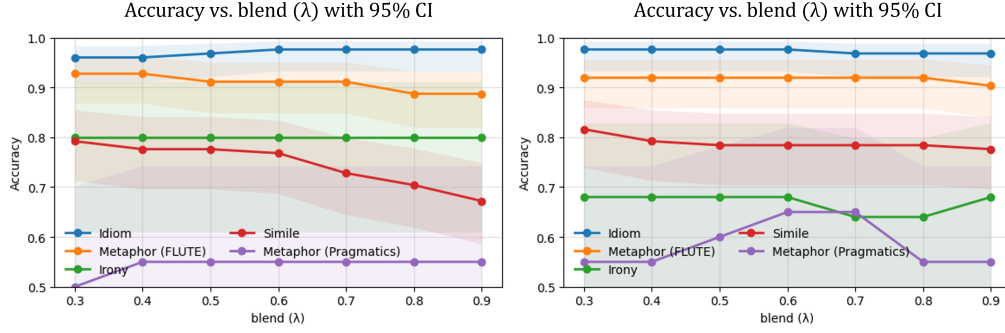


Figure 4: Each curve plots accuracy (points) with 95% confidence intervals (shaded bands) for U_{blend} as a function of the weighting coefficient λ . We vary λ from 0.3 to 0.9. The left panel reports RoBERTa results and the right panel reports DeBERTa results.

Figure 8 illustrates an example from the Pragmatics Irony dataset. In the scenario, the sentence “You are so slow!” is ironic rather than a literal statement of slowness. Hence, the correct option should form early connections to tokens conveying the opposite meaning. Option 2 links only s_{29} and s_{12} (friend) to the target token o_0 (John), with minimal contribution from other tokens. Option 3 likewise begins with links centered on o_0 and adds some connections, yet the core scenario token s_{23} (slow) remains disconnected through the third threshold. By contrast, Options 1 and 4 link s_{23} to option tokens at early thresholds. However, Option 1 leaves the meaning-determining negation token o_2 (not) unlinked until the final threshold, whereas Option 4 connects the representative phrase $o_3 o_4$ (outstanding runner) to s_{23} by the second threshold. Consequently, Option 4, capturing the ironic meaning more faithfully, achieves the lowest SATS value. A notable point is that, in Options 1 and 4, the first connection in the early thresholds does not occur with o_0 but with tokens that embody the core meaning of each option. This early activation pattern indicates that SATS prioritizes contextually coherent scenario–option linkages over mere lexical overlap.

SATS also exhibits failure cases. Figure 9 applies SATS to the same example as Figure 1. The core scenario tokens are 100 and polyester, corresponding to s_{21} and s_{23} . Although connections between these tokens and the option tokens are observed, a recurring pattern emerges: except for Option 4, the earliest threshold preferentially links to the final token of the option. That terminal token is the punctuation mark (.), which has no semantic relevance to the scenario. Consequently, in this example SATS fails to capture the implied meaning of 100 and polyester which mean “unnaturalness”, and biased by the early link to punctuation, incorrectly predicts Option 2 as correct.

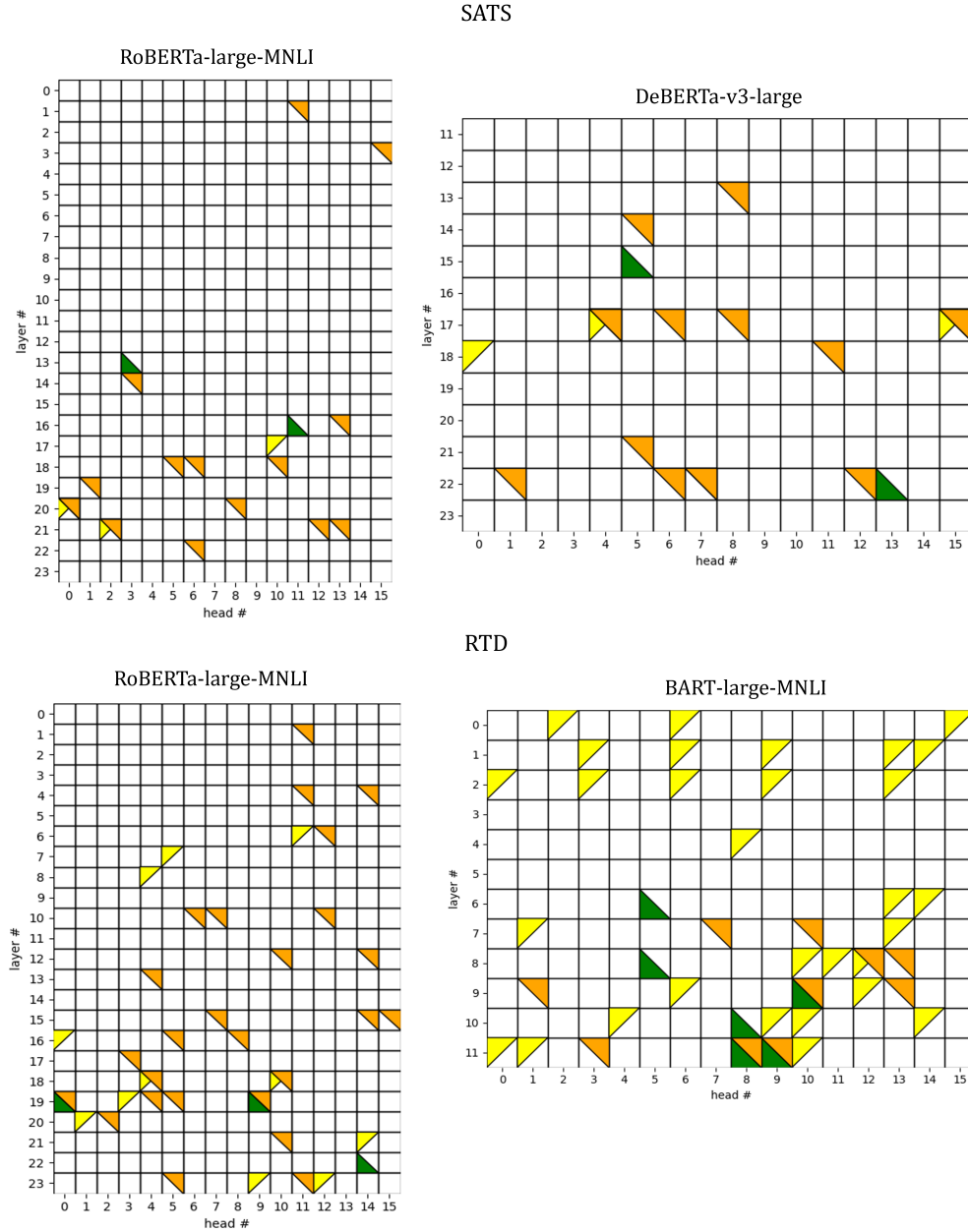


Figure 5: (Top) Heads selected by RoBERTa-large-MNLI and DeBERTa-v3-large for the best cases on each dataset in SATS. Yellow : FLUTE, Green : Pragmatics, Orange : Top 3. (Below) Heads selected by BART-large-MNLI and RoBERTa-large-MNLI for the RTD head ensemble. Yellow : Simile, Green : Metaphor, Orange : Idiom

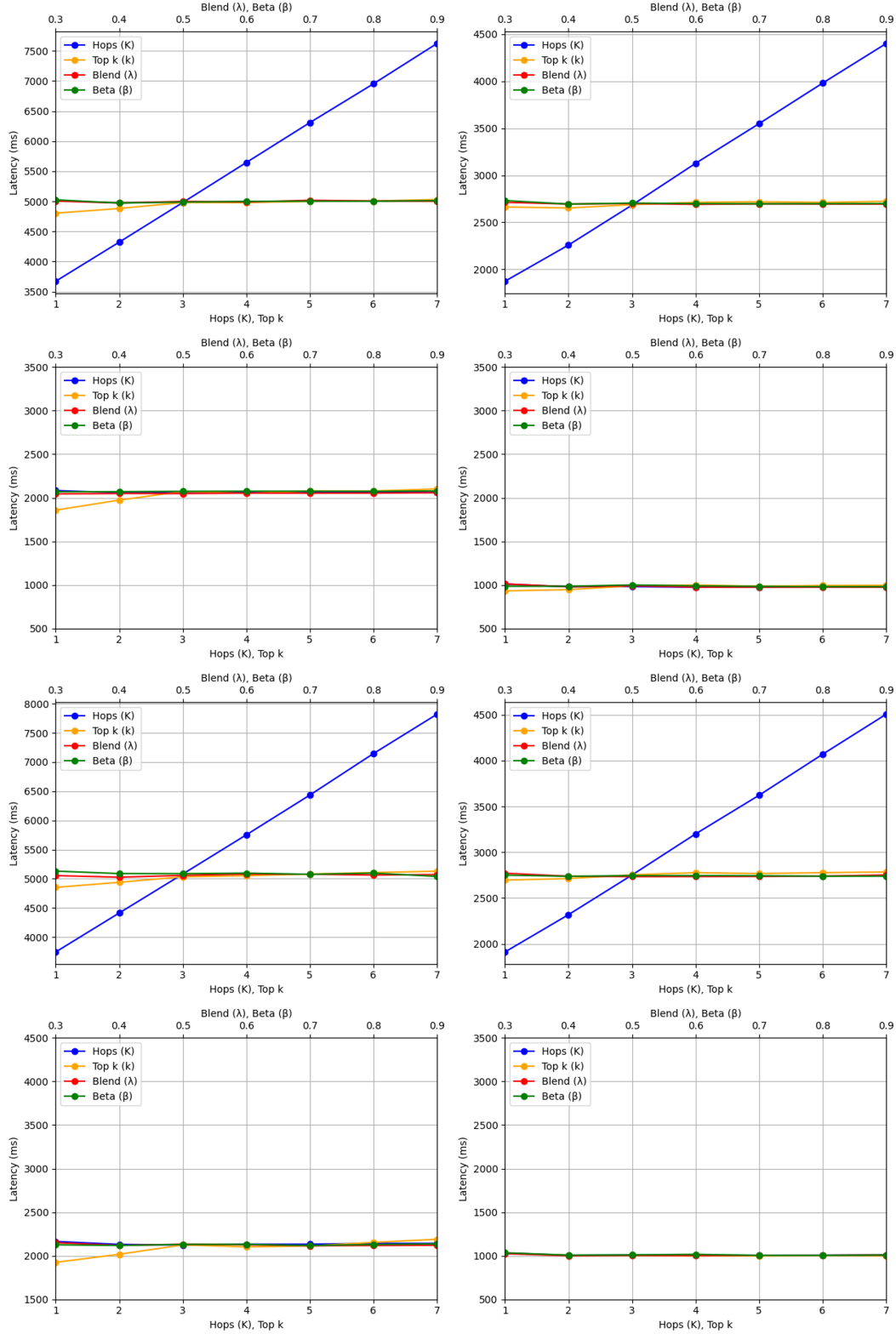


Figure 6: The top four panels show results for DeBERTa-v3-large, and the bottom four panels show results for RoBERTa-large-MNLI. The left panels correspond to the FLUTE Idiom dataset, and the right panels to the Pragmatics Metaphor dataset. For each model, the second row reports latency with multi-hop disabled. All other hyperparameters are held at their default values.

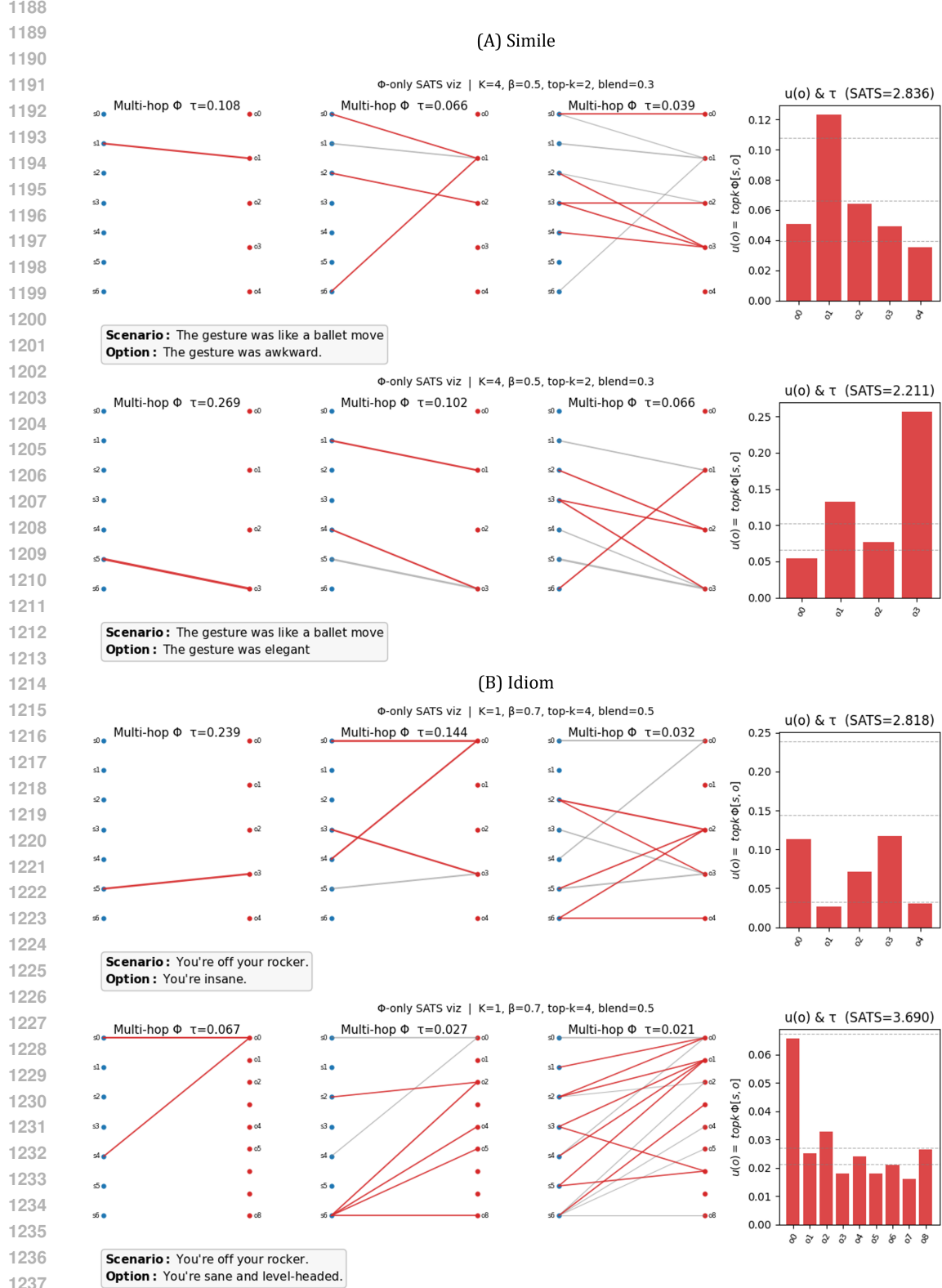


Figure 7: Example using DeBERTa on the FLUTE dataset. (A): Simile, (B): Idiom. Multi-hop Φ visualization (left) and SATS scores (right).

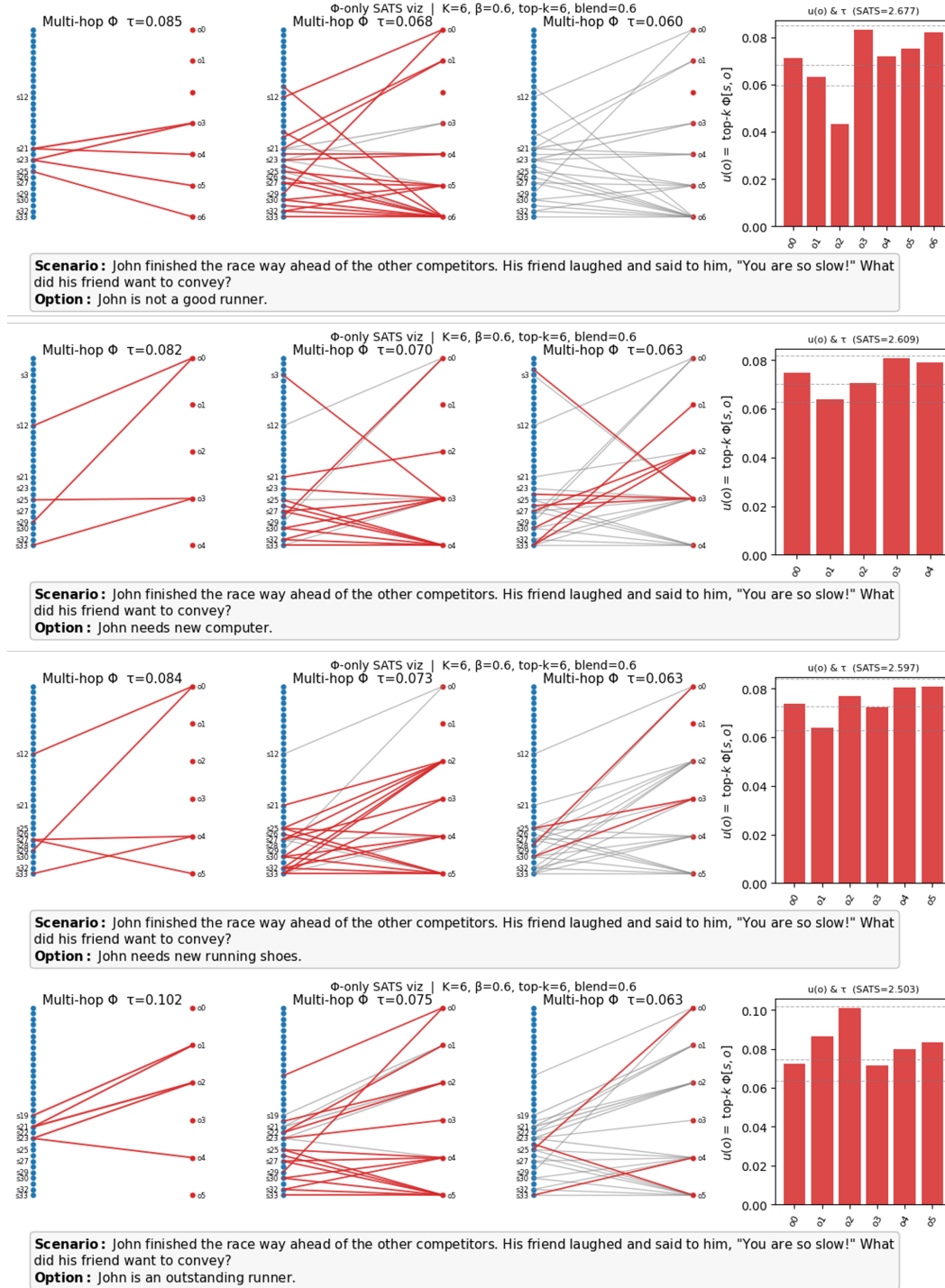


Figure 8: Results on the Pragmatics Irony dataset with DeBERTa. Multi-hop Φ visualization (left) and SATS scores (right).

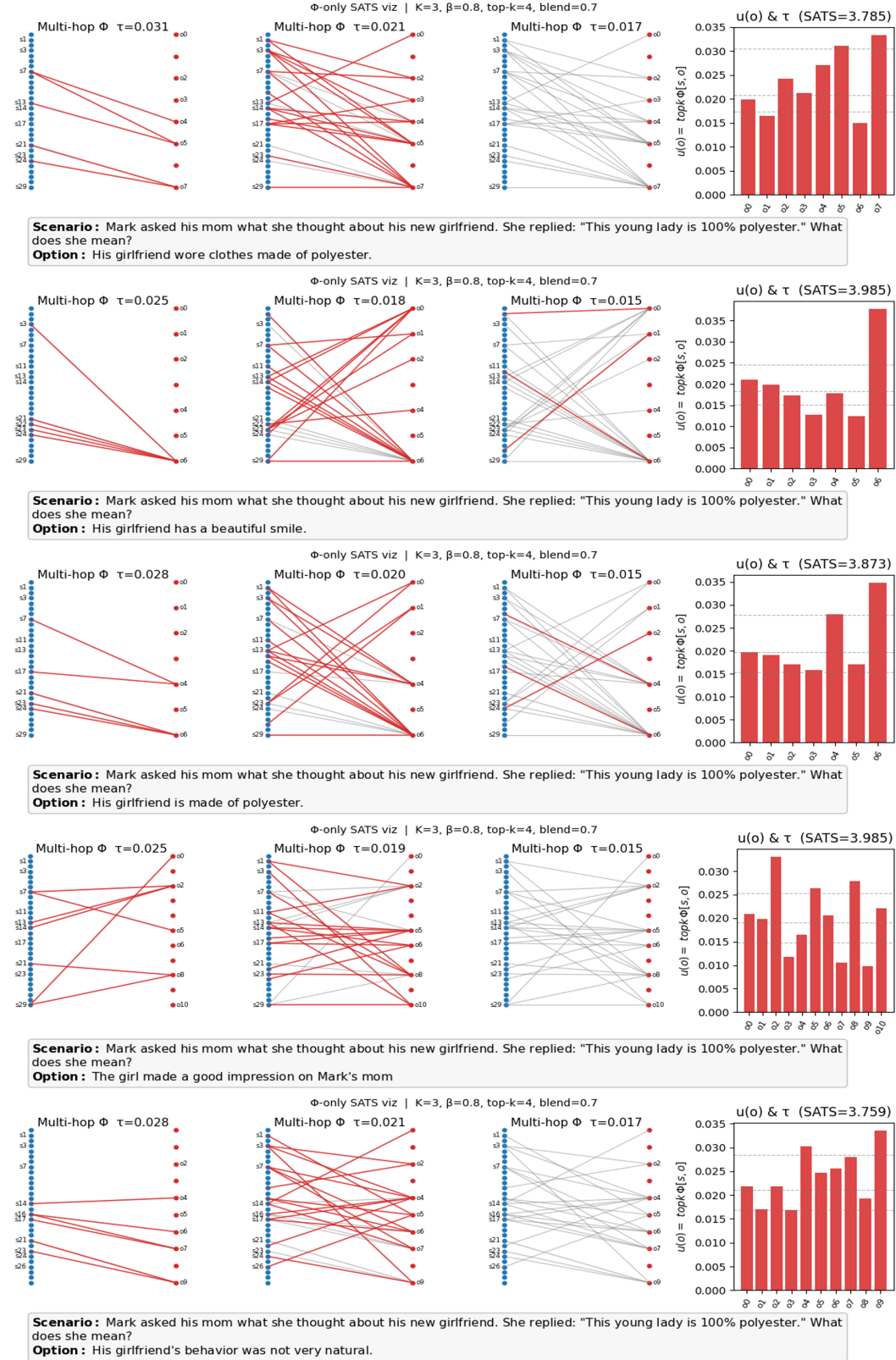


Figure 9: Same example with Figure 1 by DeBERTa: Multi-hop Φ visualization (left) and SATS scores (right).