Recurrent Stacking of Layers for Compact Neural Machine Translation Models: Supplementary Material

Raj Dabre **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
firstname.lastname@nict.go.jp

Heat-maps of Attention and Entropies

In the following pages, we give example visualizations of full sentence (sub-word segmented) encoder self-attention, decoder self-attention and encoder-decoder cross-attention for three different Japanese sentences taken from the GCP Japanese-to-English configuration listed in Table 1. Note that the heat-maps include the sub-word segmented version of the sentence, and that the translations decoded by the vanilla NMT and the RS-NMT models are not necessarily identical.

It can be seen that the observations we mentioned in the main content of the paper hold true for all these examples. Note that these are not cherry-picked examples and are completely randomly chosen.

a) お 釣り の ２００ 円 です 。
Here is your 200 yen change.

Figure #	Vanilla	RS-NMT	Entropies
Encoder self-attention	1	2	3
Decoder self-attention	4	5	6
Cross-attention	7	8	9

b) 健康保険 証 も 持っ て い ます 。
(I also have a health insurance card.)

Figure #	Vanilla	RS-NMT	Entropies
Encoder self-attention	10	11	12
Decoder self-attention	13	14	15
Cross-attention	16	17	18

c) どのように 体調 が 優れ ない の です か ?
(How don't you feel well?)

Figure #	Vanilla	RS-NMT	Entropies
Encoder self-attention	19	20	21
Decoder self-attention	22	23	24
Cross-attention	25	26	27

Table 1: Example Japanese sentences taken from the GCP Japanese-to-English configuration, accompanied by manual English gloss. The "Vanilla" columns and "RS-NMT" columns show the Figure numbers for the 6-layer vanilla NMT models and for the 6-layer RS-NMT models, respectively. The "Entropies" columns present the Figure numbers that compare the layer×head attention entropies of the vanilla NMT and RS-NMT models.

Modifications to The Transformer Code

In order to enable RS in the Transformer, a simple modification is required. Although we used version 1.6, this modification can be used in version 1.8 as well. Note that version 1.8 has an implementation for Universal Transformer which is a lot more complex than our RS-NMT model. The modifications are as follows:

Code Type : Original
Line 1213 – with tf.variable_scope("layer_%d" % layer):
Line 1299 – with tf.variable_scope(layer_name):

Code Type : Modified
Line 1213 – with tf.variable_scope("layer_0", reuse = True if layer != 0 else False):
Line 1299 – with tf.variable_scope("layer_0", reuse = True if layer != 0 else False):

Note that the "layer" in the above code refers to the for-loop variable which counts the number of layers to be stacked.

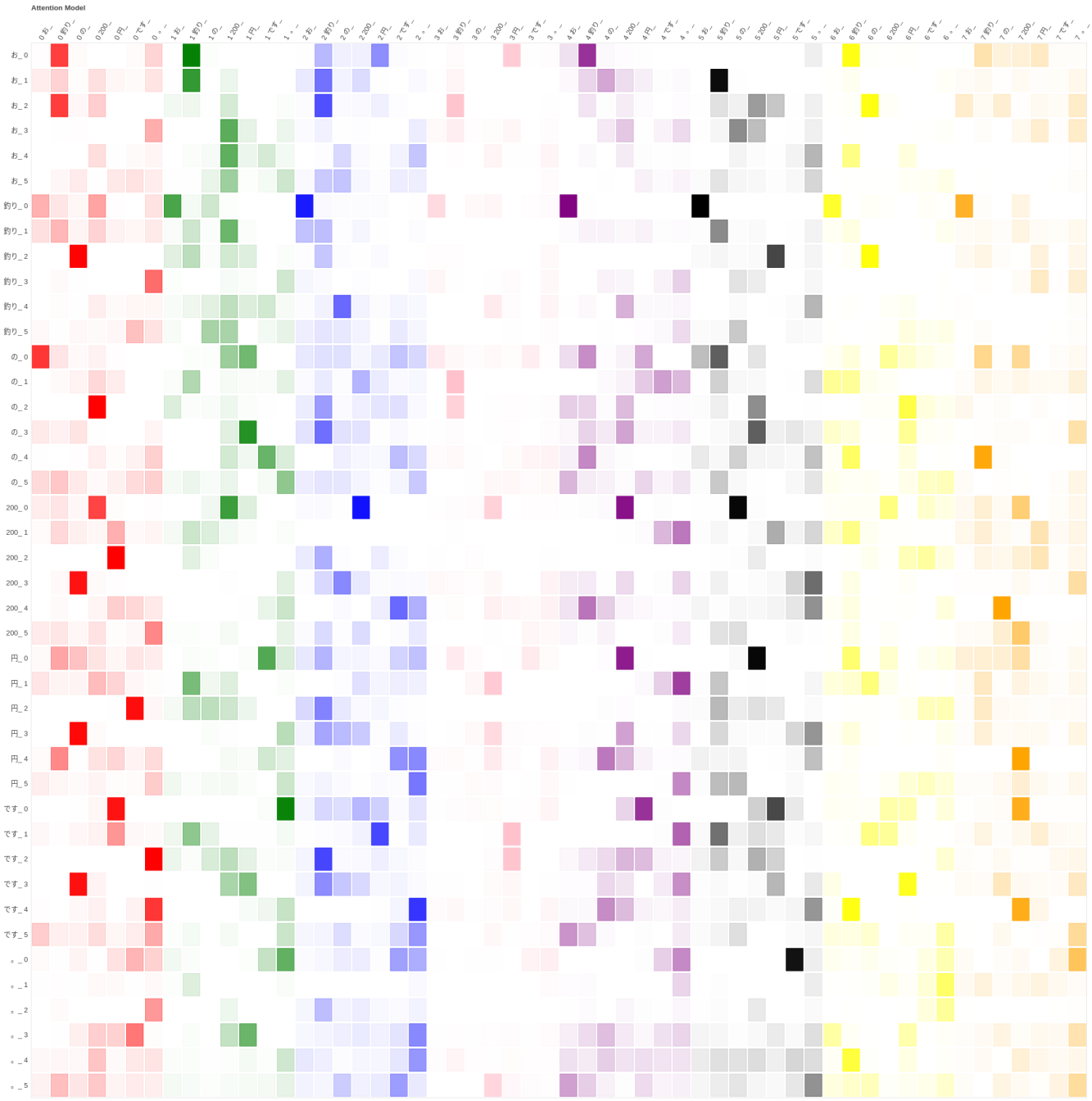

Figure 1: Vanilla 6-layer NMT model's attention heat-map for the encoder self-attention for the input sentence: "お 釣り の 2 0 0 円 です 。" (Here is your 200 yen change.).

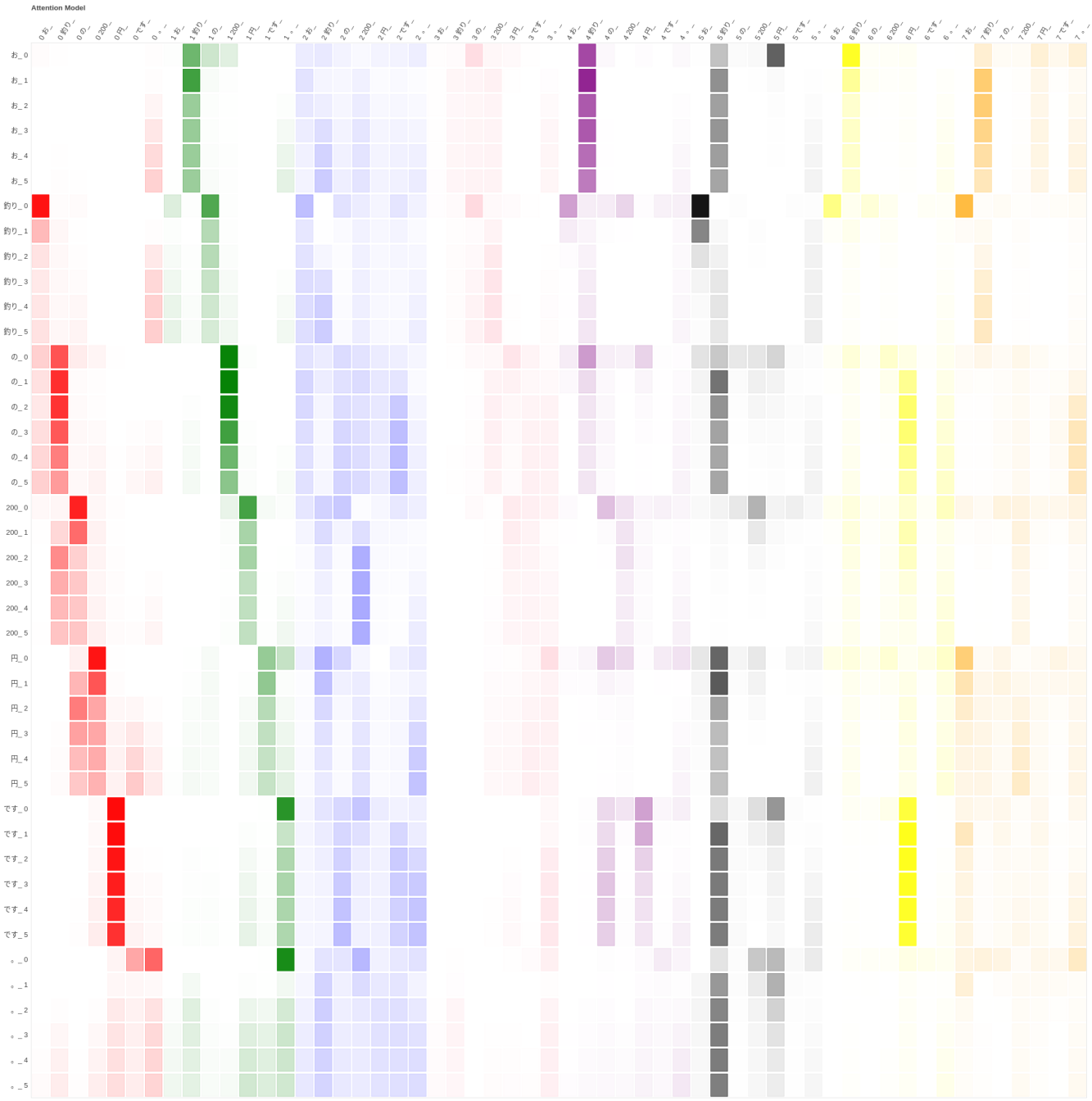

Figure 2: 6-layer RS-NMT model's attention heat-map for the encoder self-attention for the input sentence is "お 釣り の ２ ０ ０ 円 です 。" (Here is your 200 yen change.).

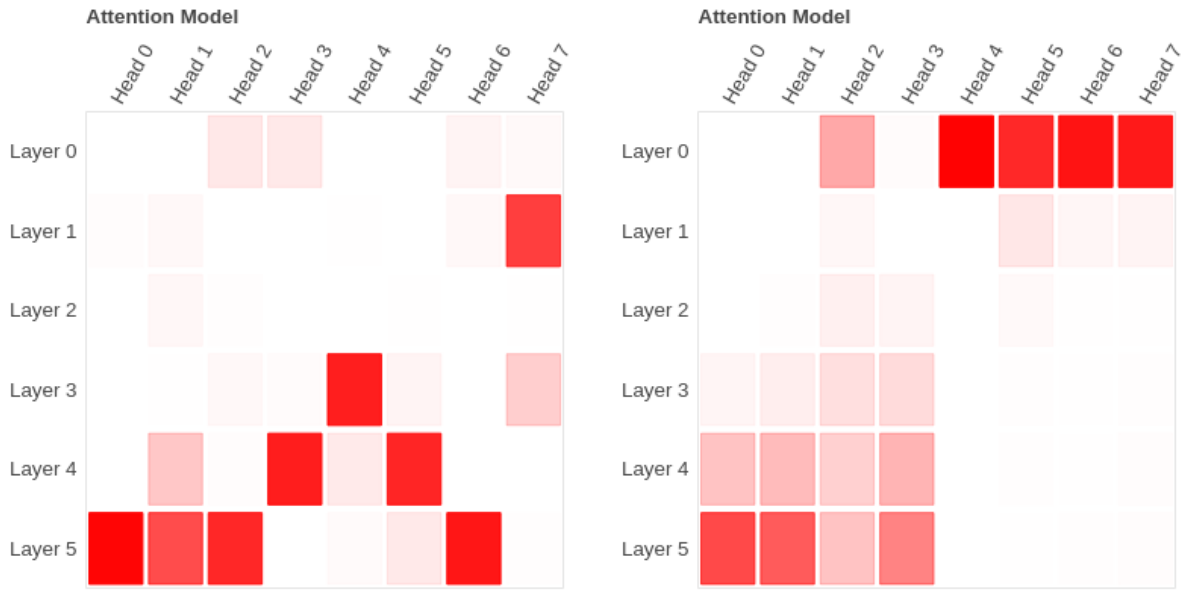

Figure 3: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder self-attentions for the input sentence "お 釣り の ２００ 円 です 。" (Here is your 200 yen change.).

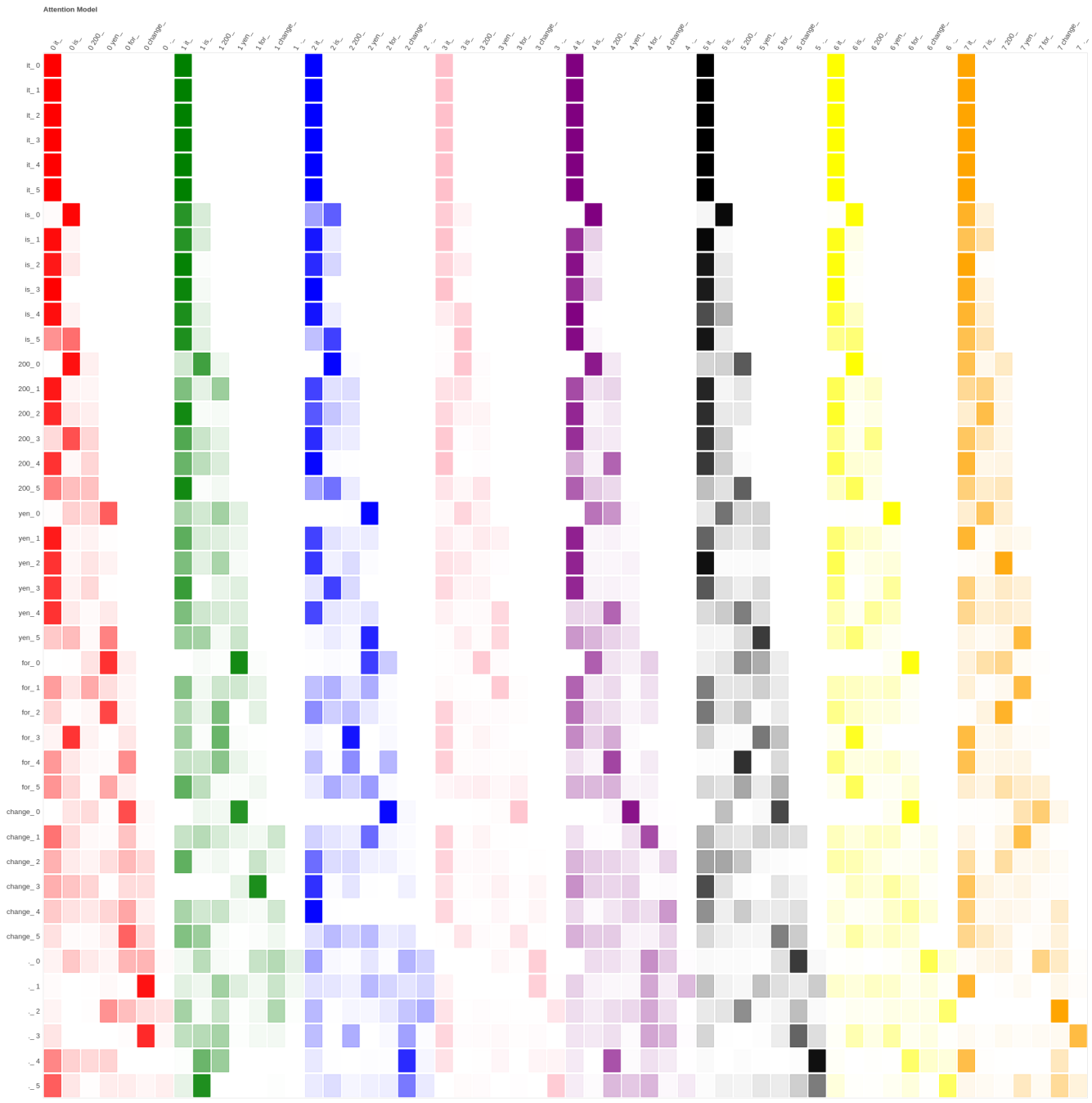

Figure 4: Vanilla 6-layer NMT model's attention heat-map for the decoder self-attention for the generated target sentence: "it is 200 yen for change". The input sentence is "お 釣り の ２００ 円 です 。" (Here is your 200 yen change.).

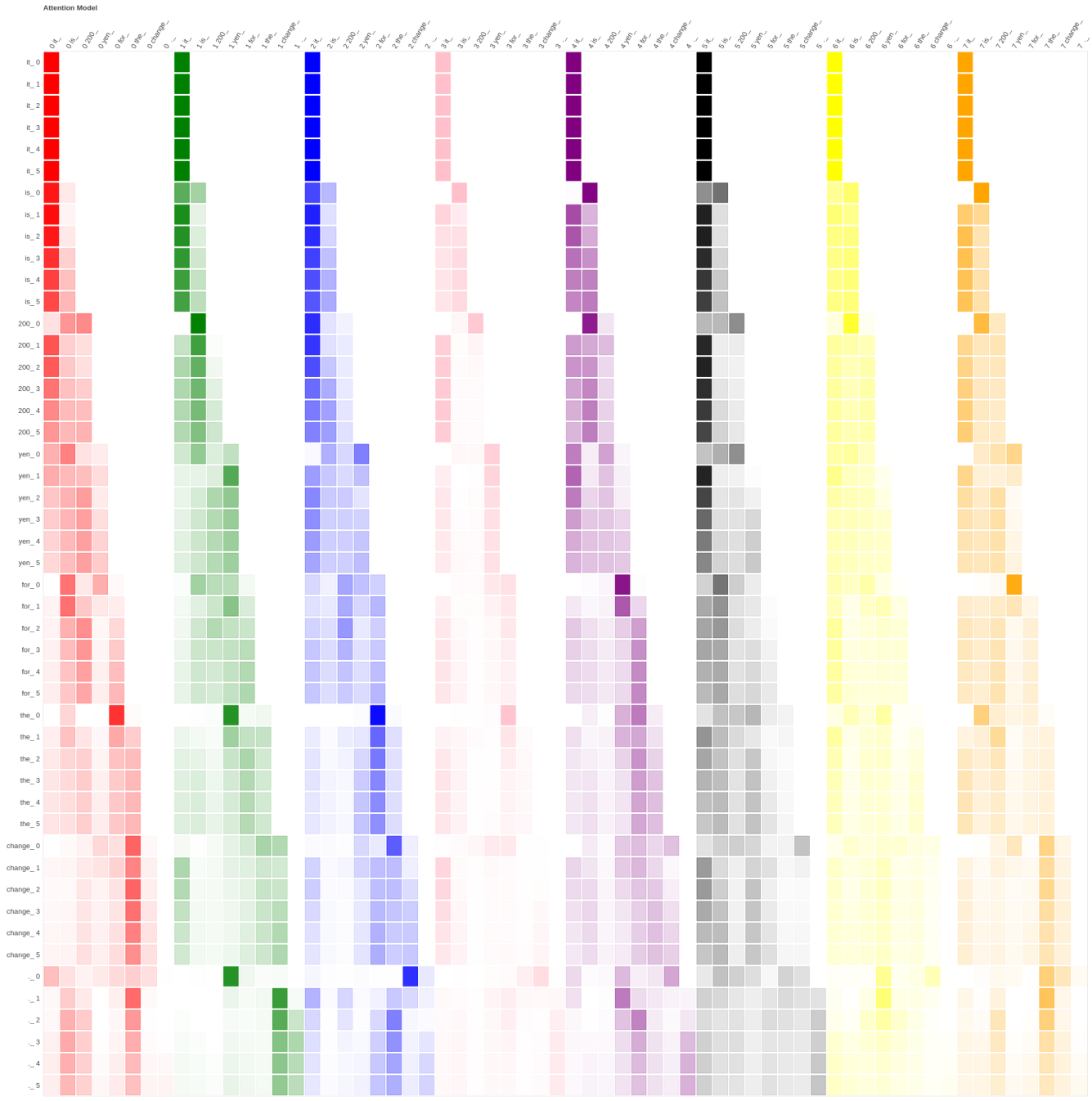

Figure 5: 6-layer RS-NMT model's attention heat-map for the decoder self-attention for the generated target sentence: "it is 200 yen for the change". The input sentence is "お釣りの２００円です。" (Here is your 200 yen change.).

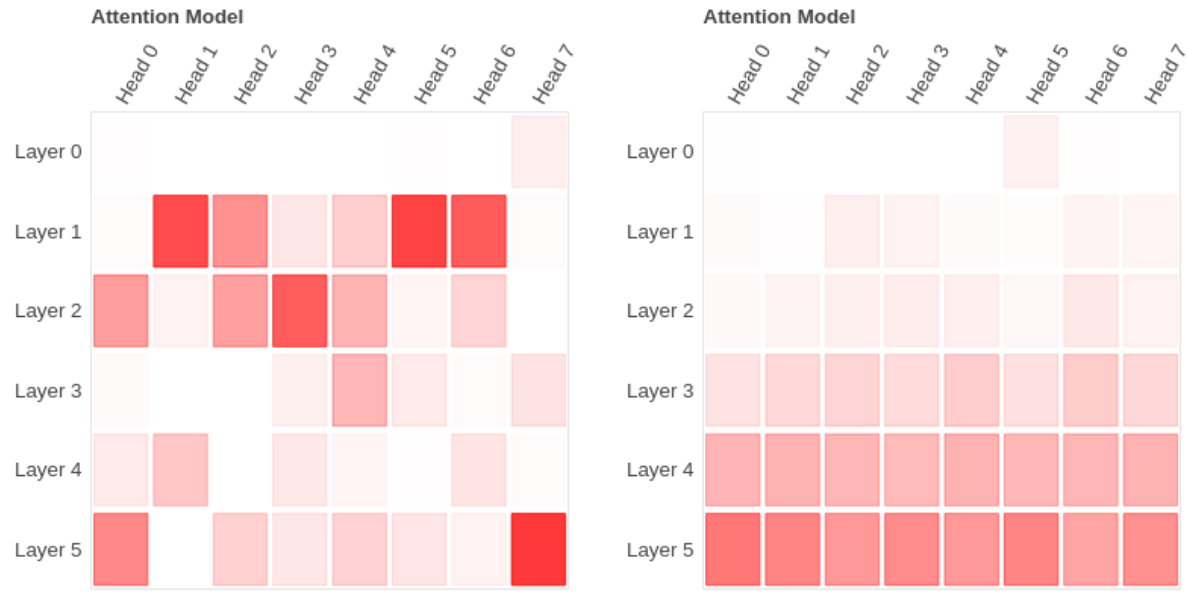

Figure 6: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) decoder self-attentions for the generated target sentences: "it is 200 yen for change" and "it is 200 yen for the change". The input sentence is "お 釣り の ２００ 円 です 。" (Here is your 200 yen change.).

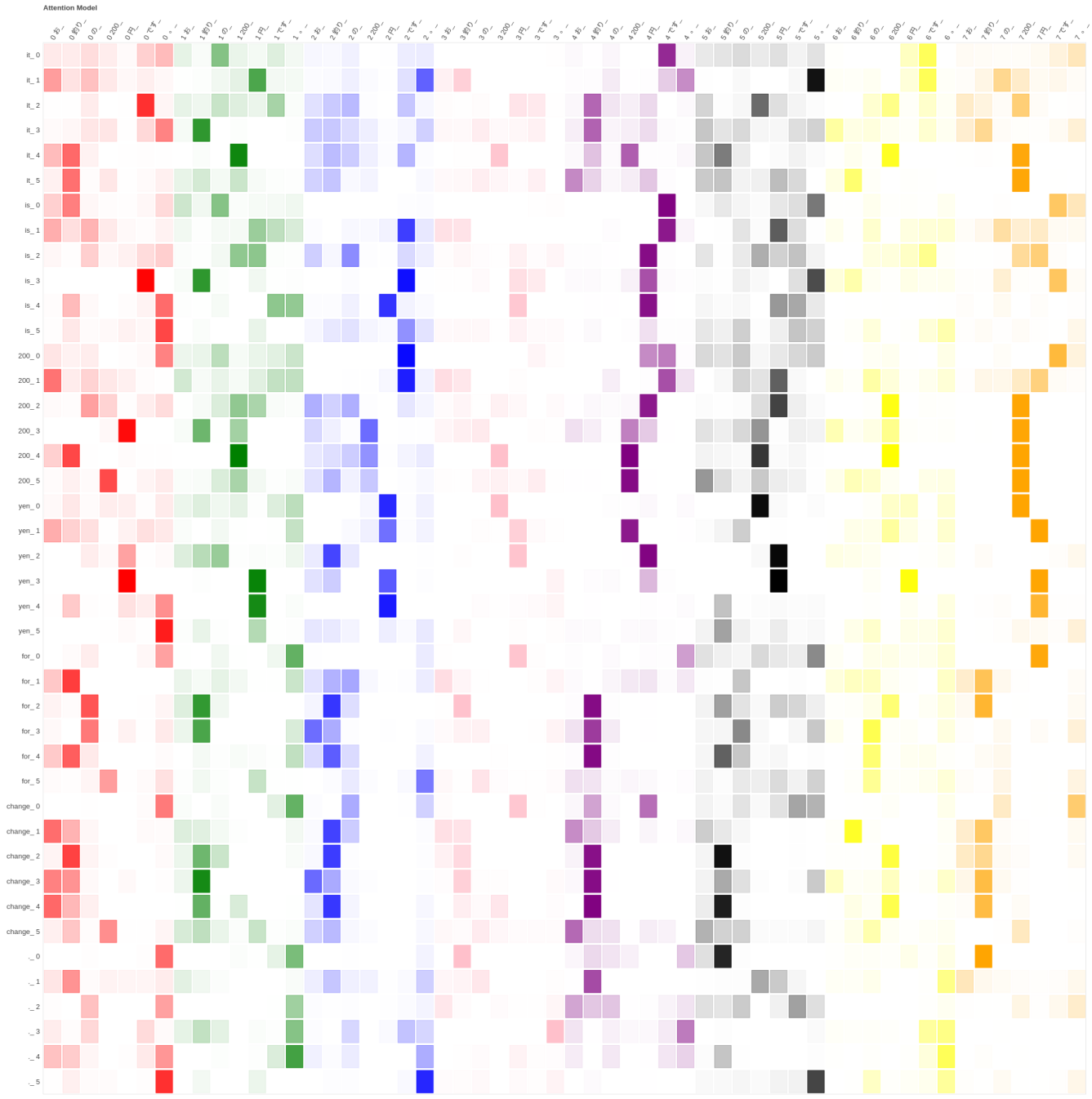

Figure 7: Vanilla 6-layer NMT model's attention heat-map for the encoder-decoder cross-attention for the generated target sentence: "it is 200 yen for change". The input sentence is "お 釣り の ２ ０ ０ 円 です 。" (Here is your 200 yen change.).

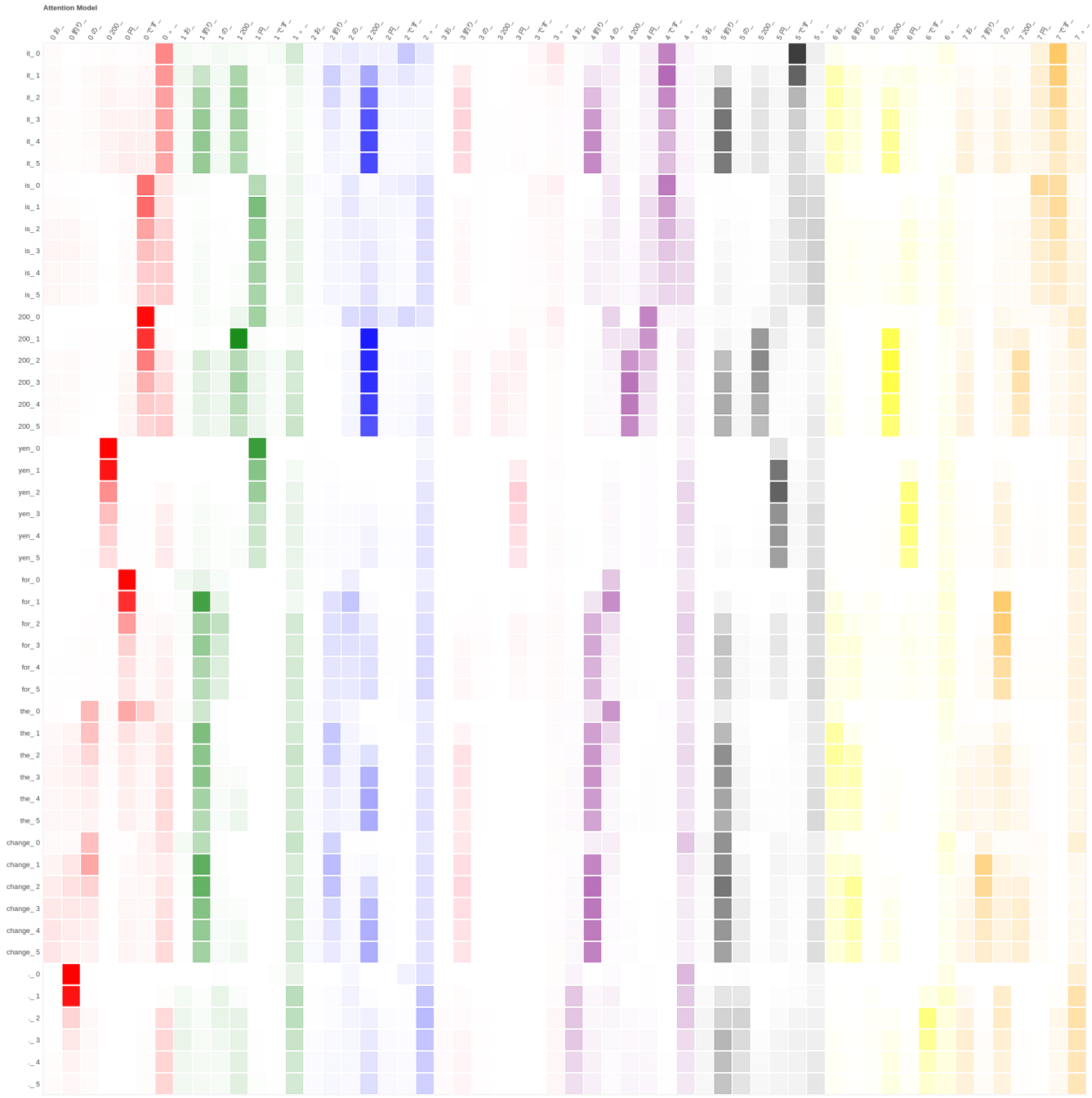

Figure 8: 6-layer RS-NMT model's attention heat-map for the encoder-decoder cross-attention for the generated target sentence: "it is 200 yen for the change". The input sentence is "お 釣り の ２００ 円 です 。" (Here is your 200 yen change.).

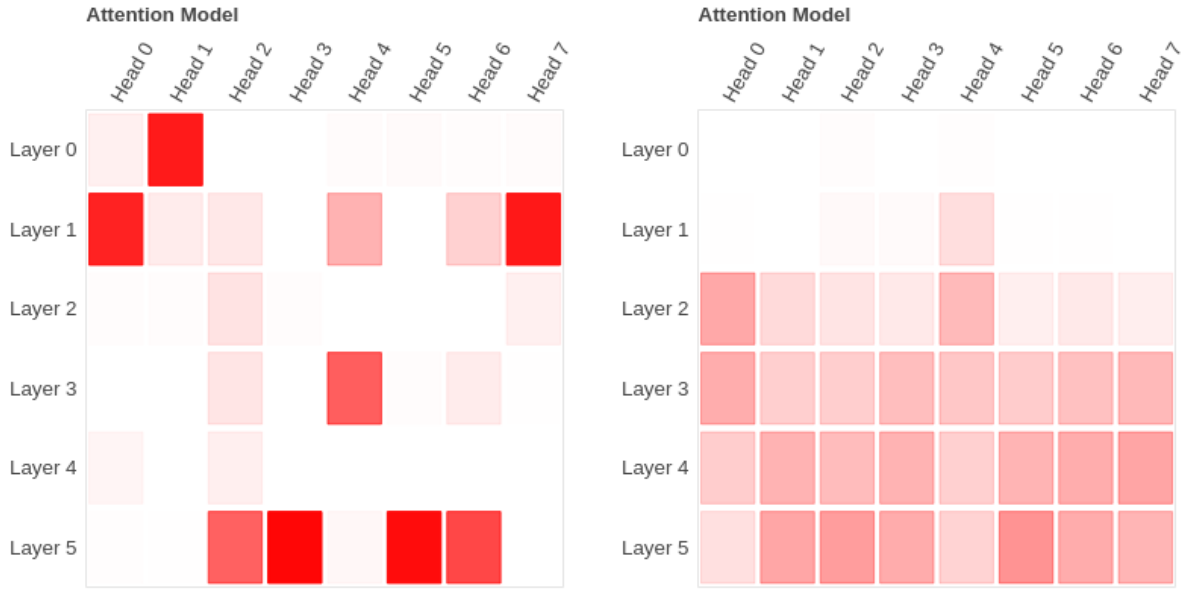

Figure 9: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder-decoder cross-attention for the generated target sentences: "it is 200 yen for change" and "it is 200 yen for the change". The input sentence is "お 釣り の ２００ 円 です 。" (Here is your 200 yen change.).

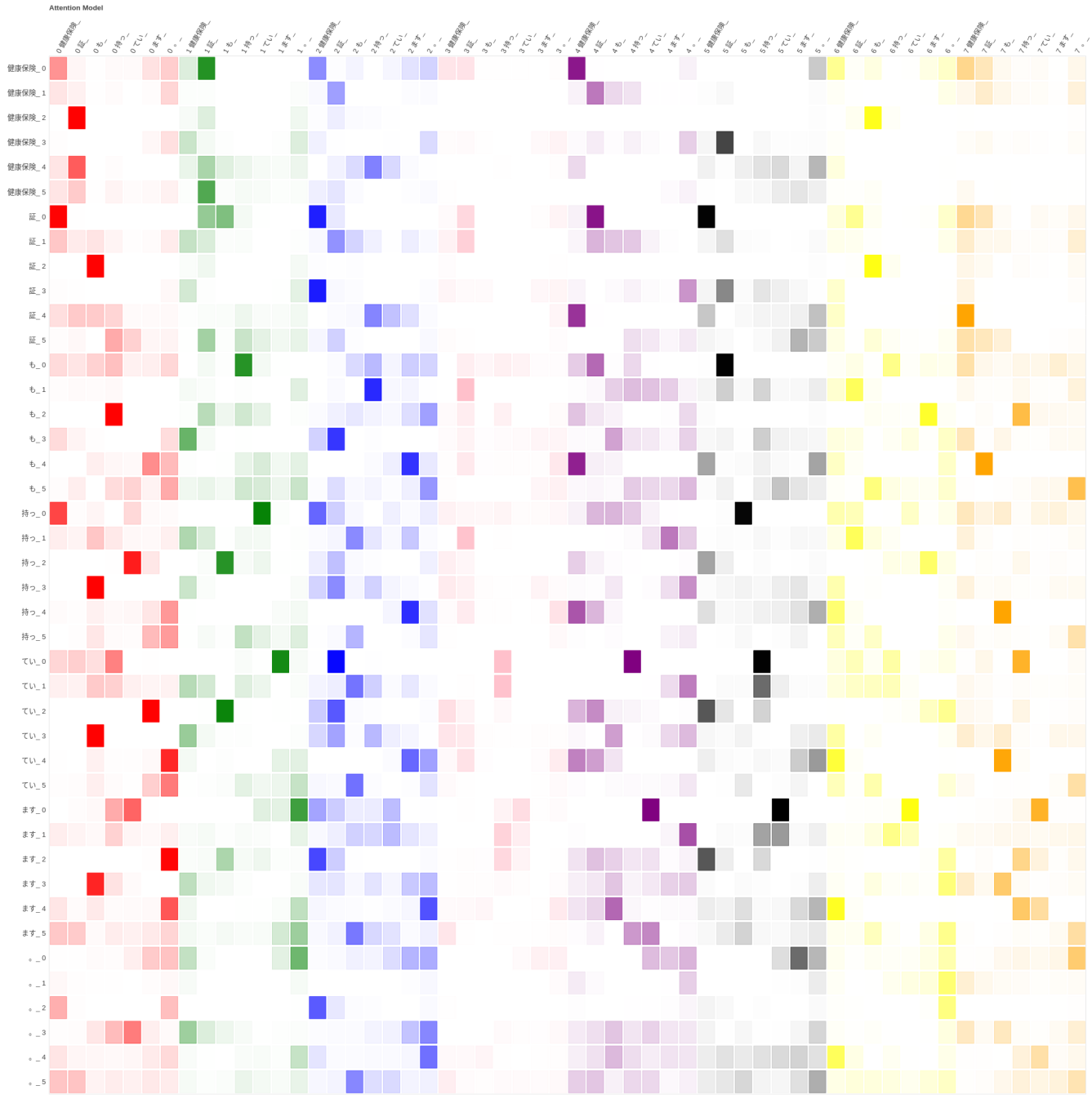

Figure 10: Vanilla 6-layer NMT model's attention heat-map for the encoder self-attention for the input sentence: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

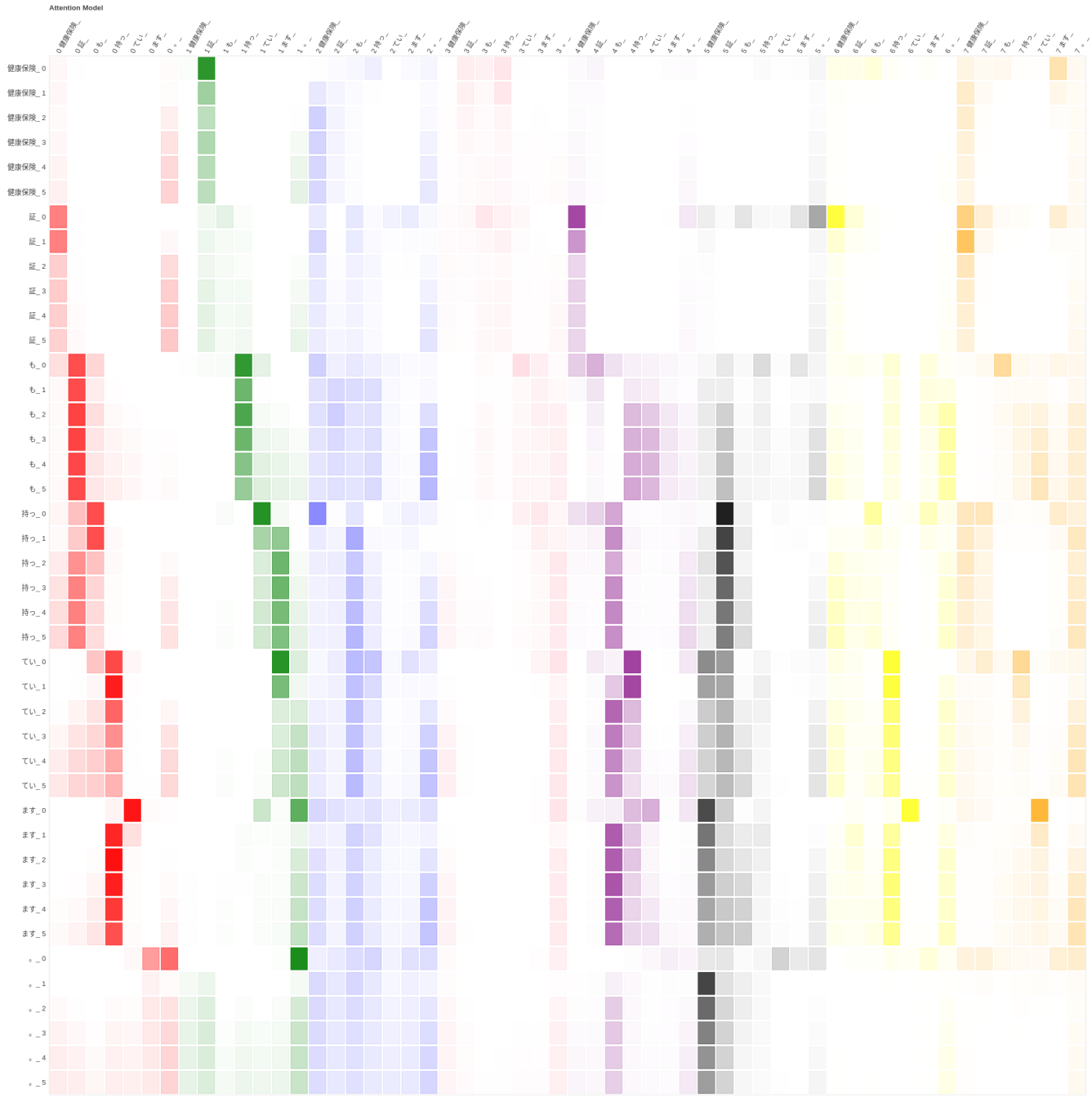

Figure 11: 6-layer RS-NMT model's attention heat-map for the encoder self-attention for the input sentence: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

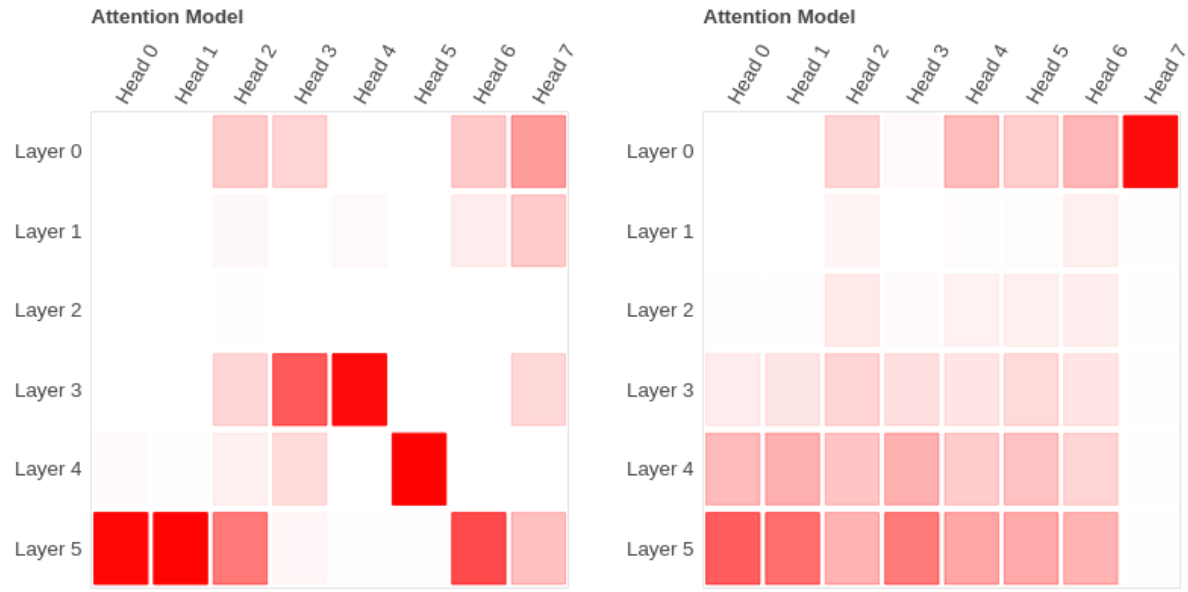

Figure 12: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder self-attention for the input sentence: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

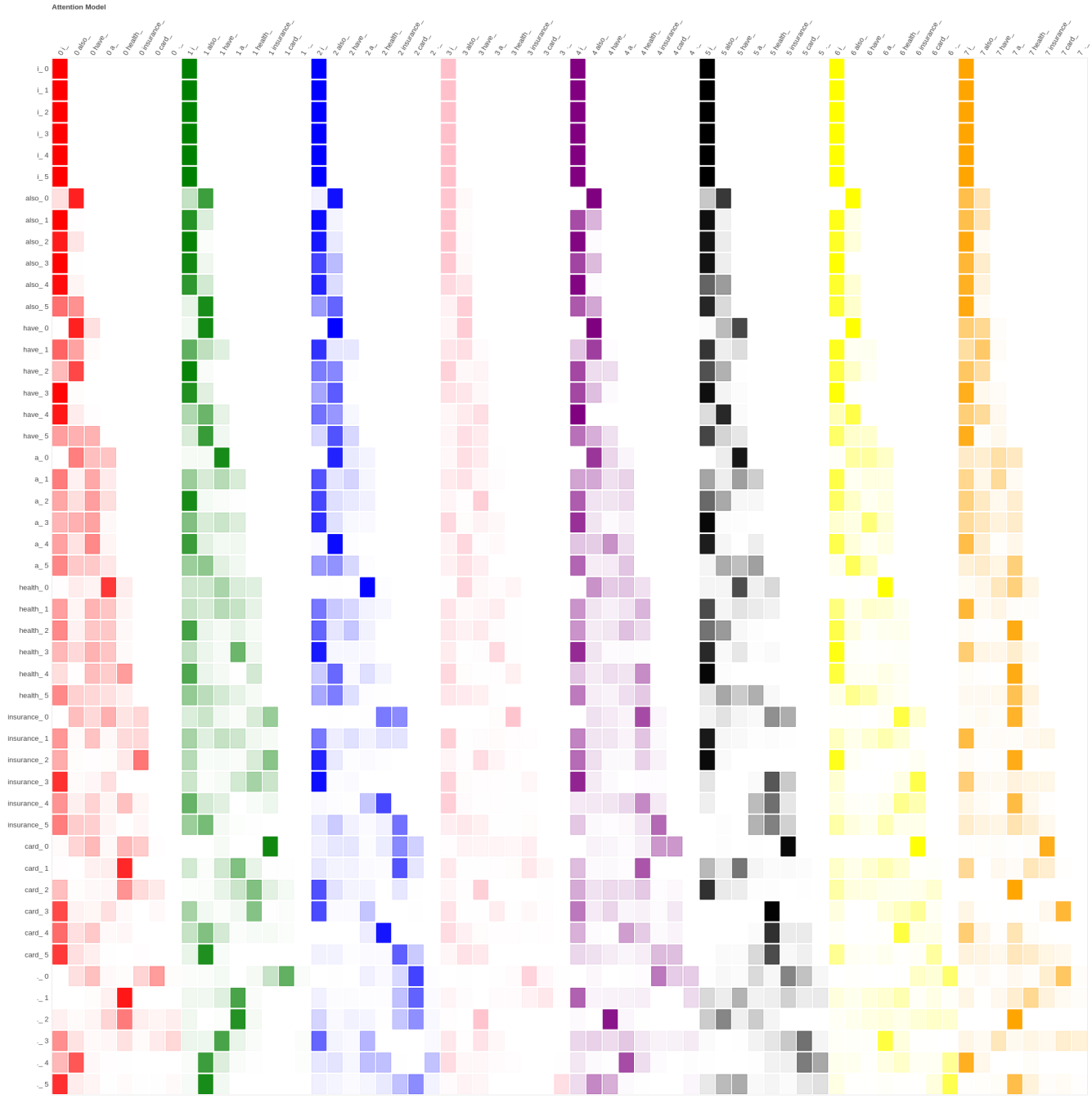

Figure 13: Vanilla 6-layer NMT model's attention heat-map for the decoder self-attention for the output sentence "i also have a health insurance card" for which the input sentence is: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

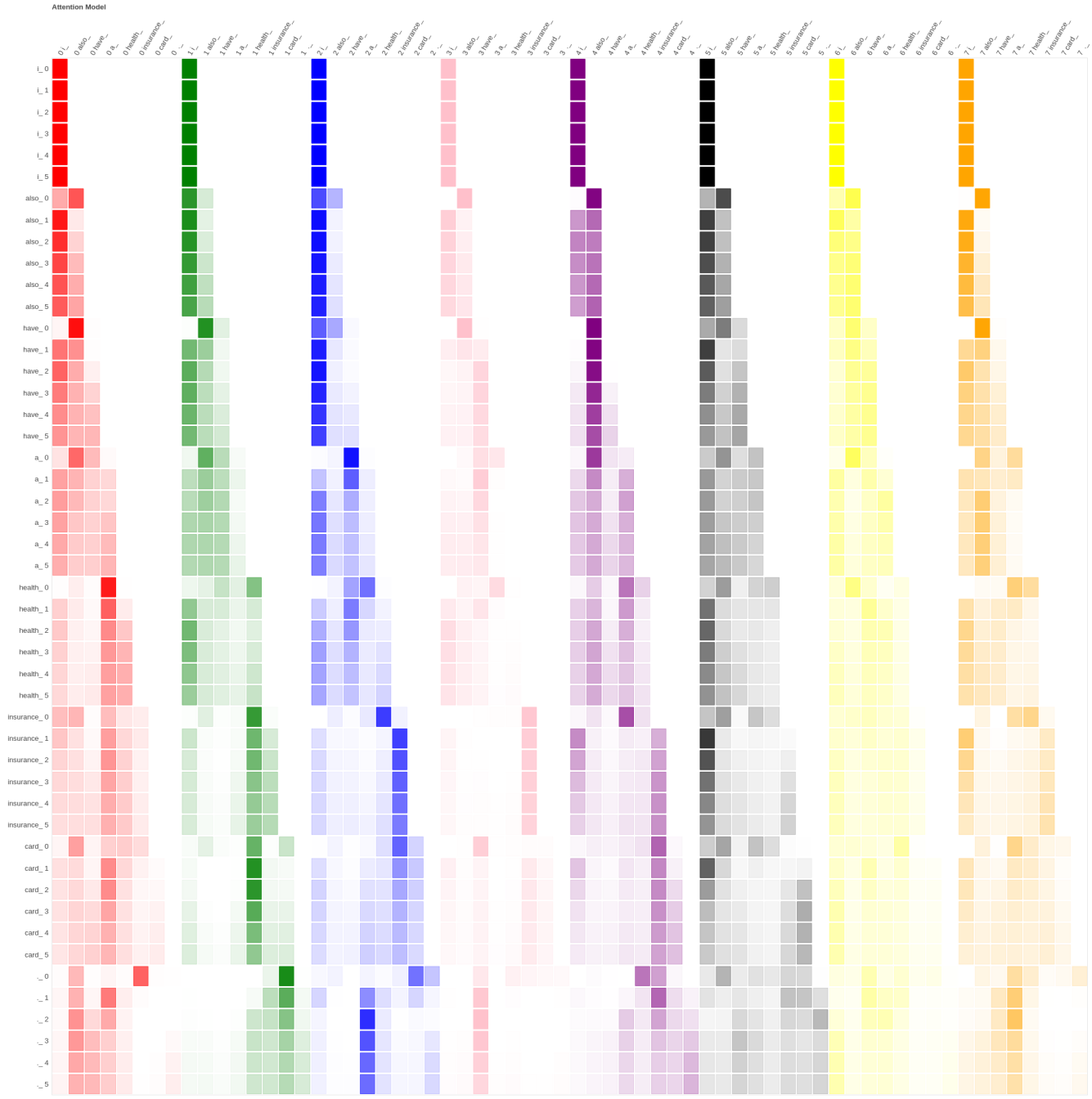

Figure 14: 6-layer RS-NMT model's attention heat-map for the decoder self-attention for the output sentence "i also have a health insurance card" for which the input sentence is: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

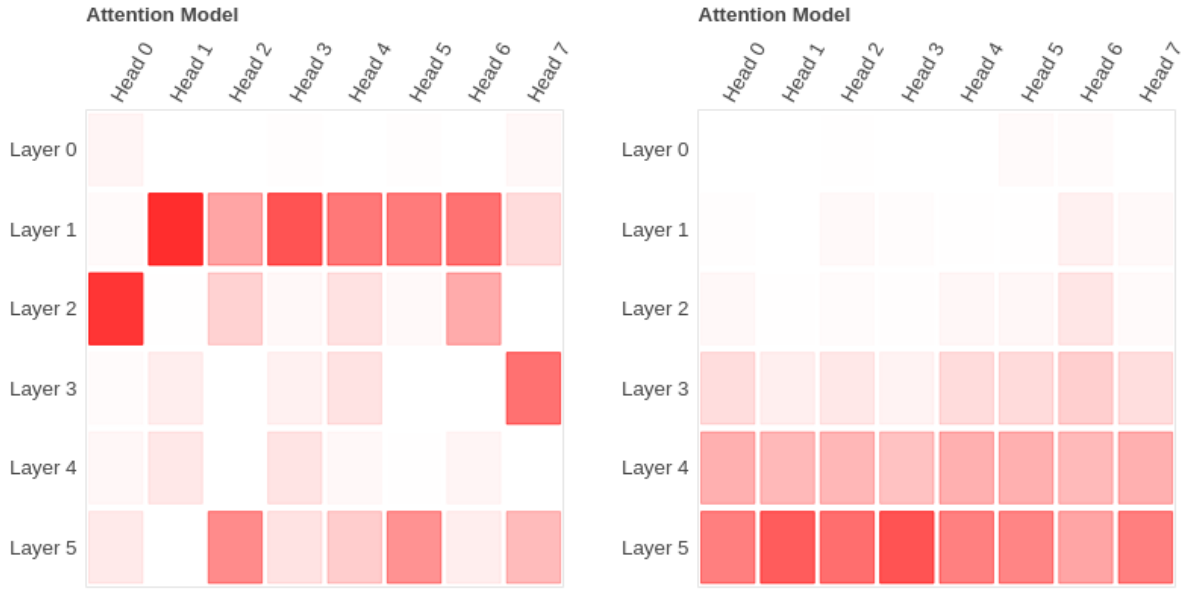

Figure 15: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) decoder self-attentions for the generated target sentences: "i also have a health insurance card" and "i also have a health insurance card" (identical). The input sentence is: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

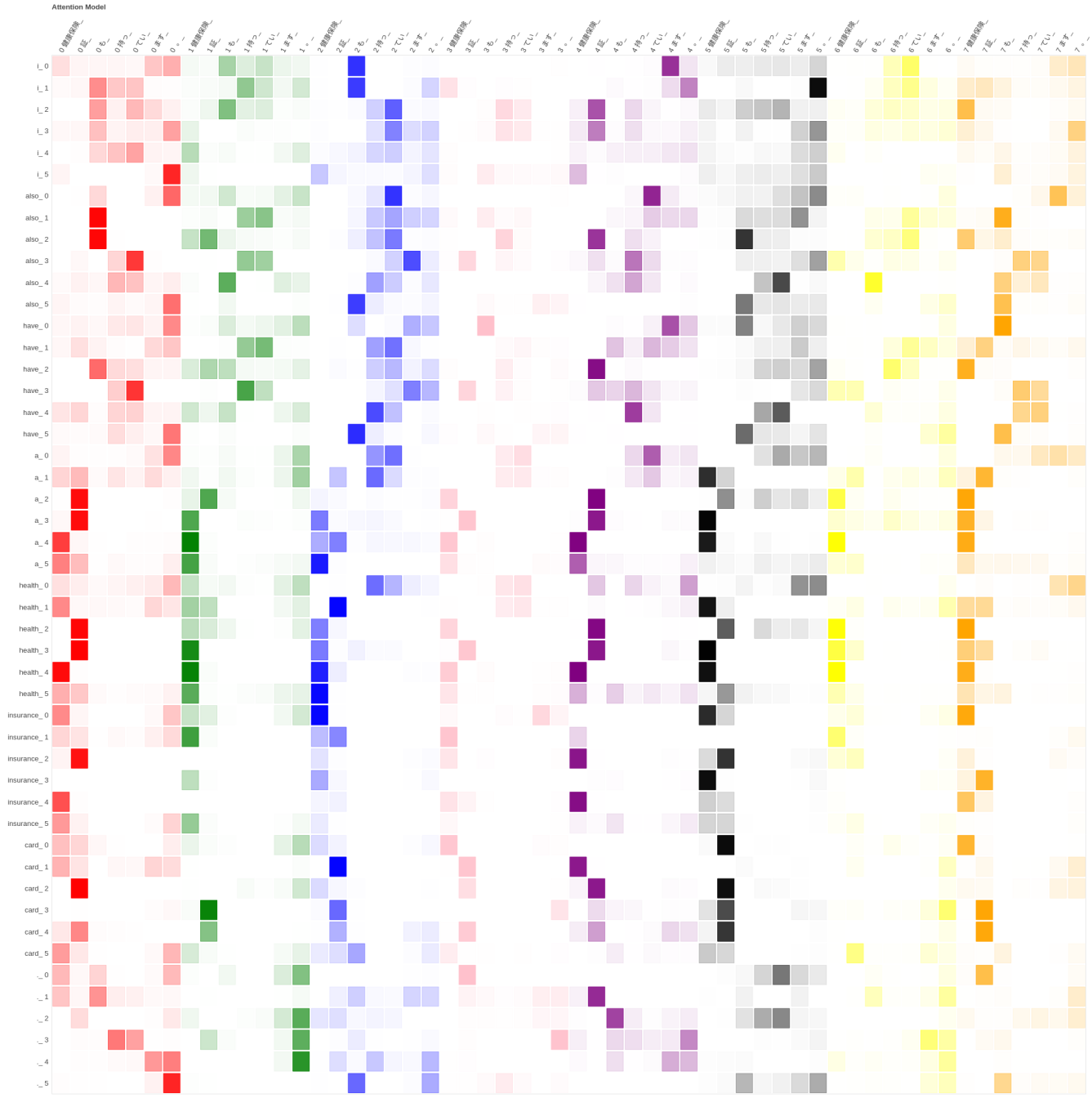

Figure 16: Vanilla 6-layer NMT model's attention heat-map for the encoder-decoder cross-attention for the output sentence "i also have a health insurance card" for which the input sentence is: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

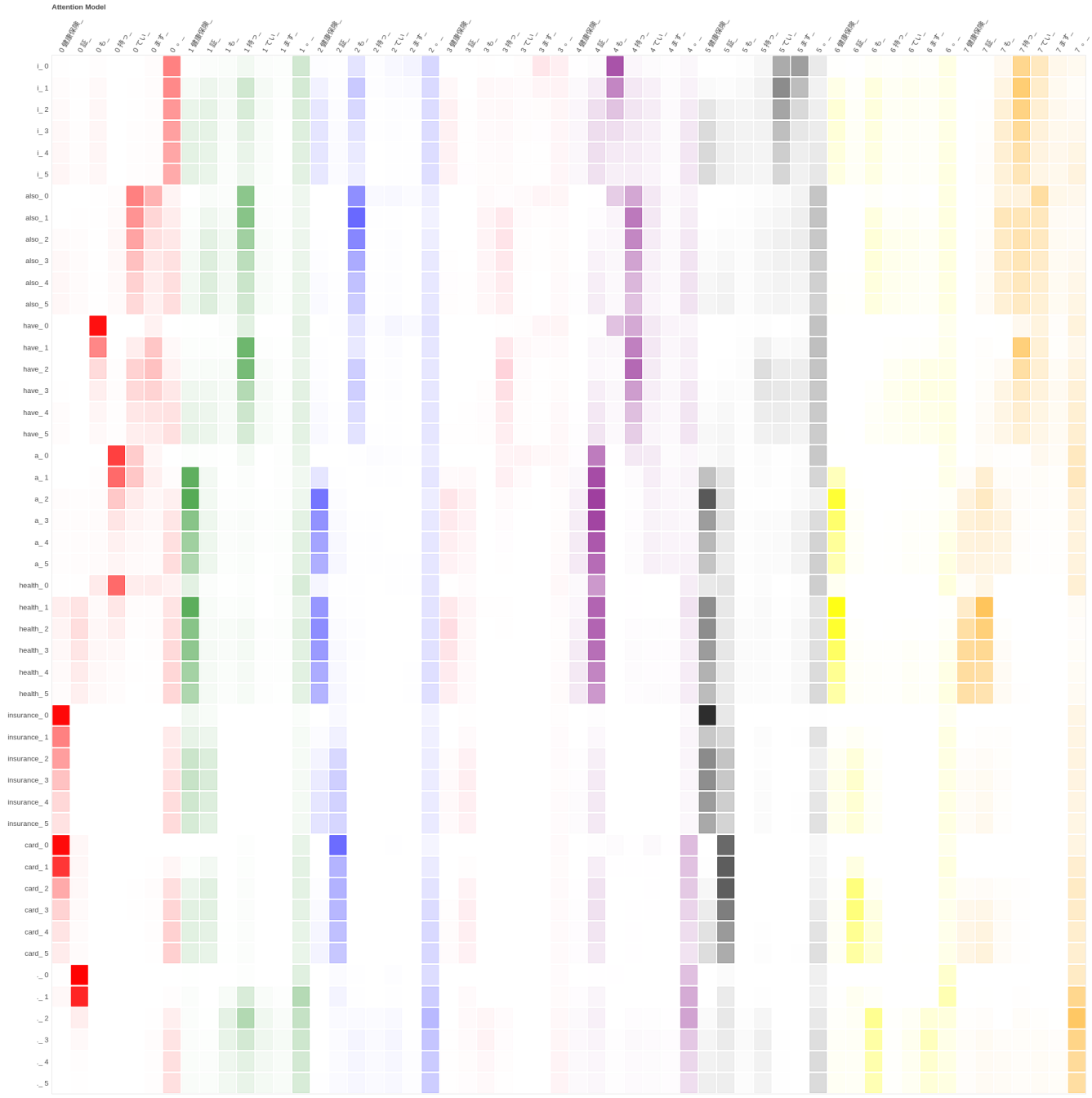

Figure 17: 6-layer RS-NMT model's attention heat-map for the encoder-decoder cross-attention for the output sentence "i also have a health insurance card" for which the input sentence is: "健康保険 証 も 持っ て い ます 。" (I also have a health insurance card.).

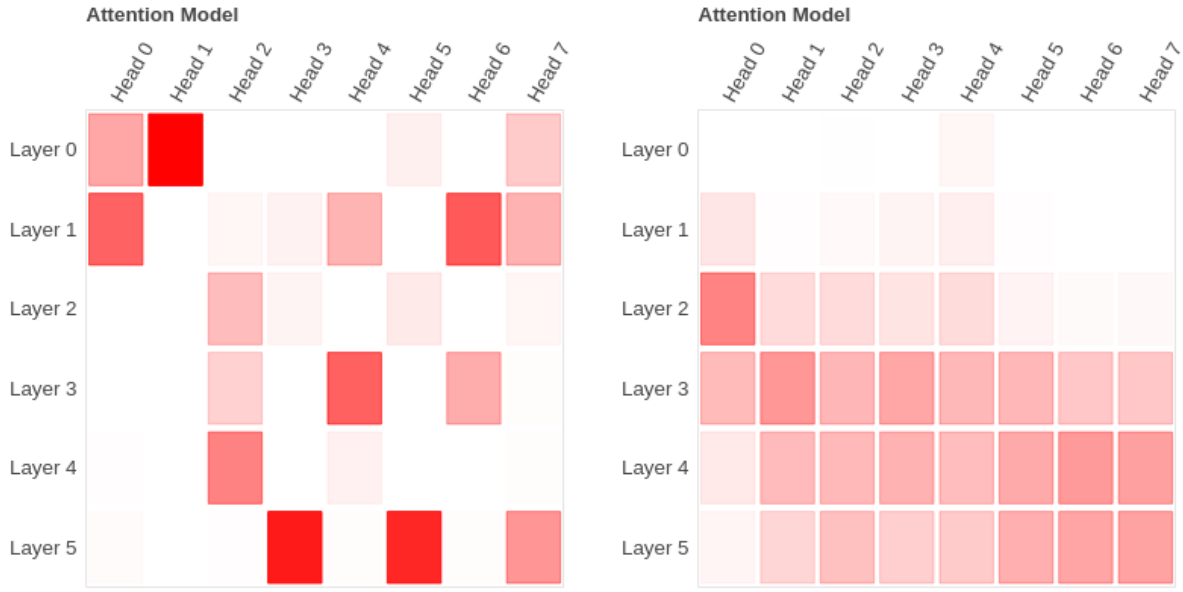

Figure 18: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder-decoder cross-attention for the output sentences "i also have a health insurance card" and "i also have a health insurance card" (identical). The input sentence is: "健康保険 証 も 持っ てい ます 。" (I also have a health insurance card.).

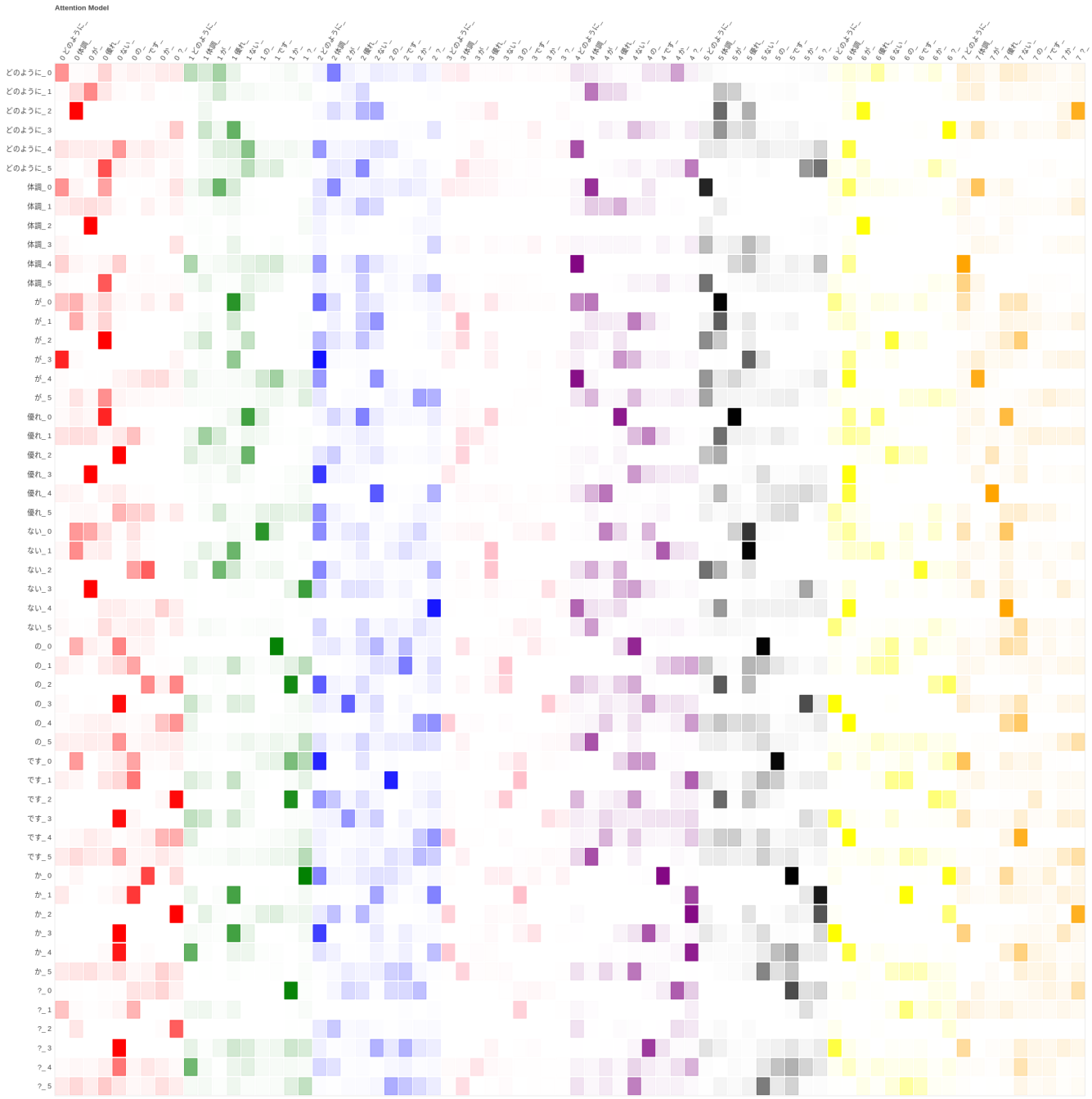

Figure 19: Vanilla 6-layer NMT model's attention heat-map for the encoder self-attention for the input sentence: "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

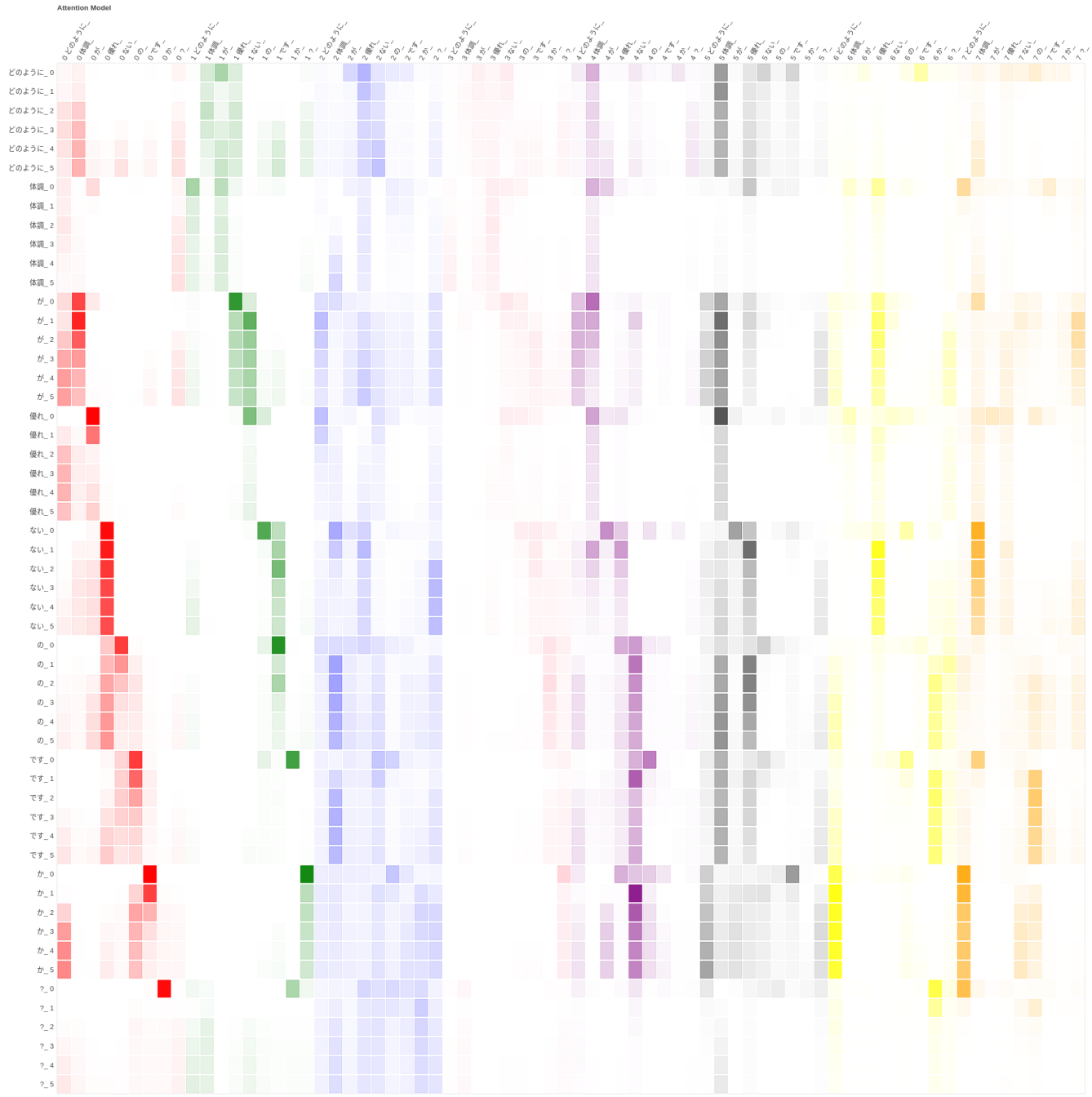

Figure 20: 6-layer RS-NMT model's attention heat-map for the encoder self-attention for the input sentence: "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

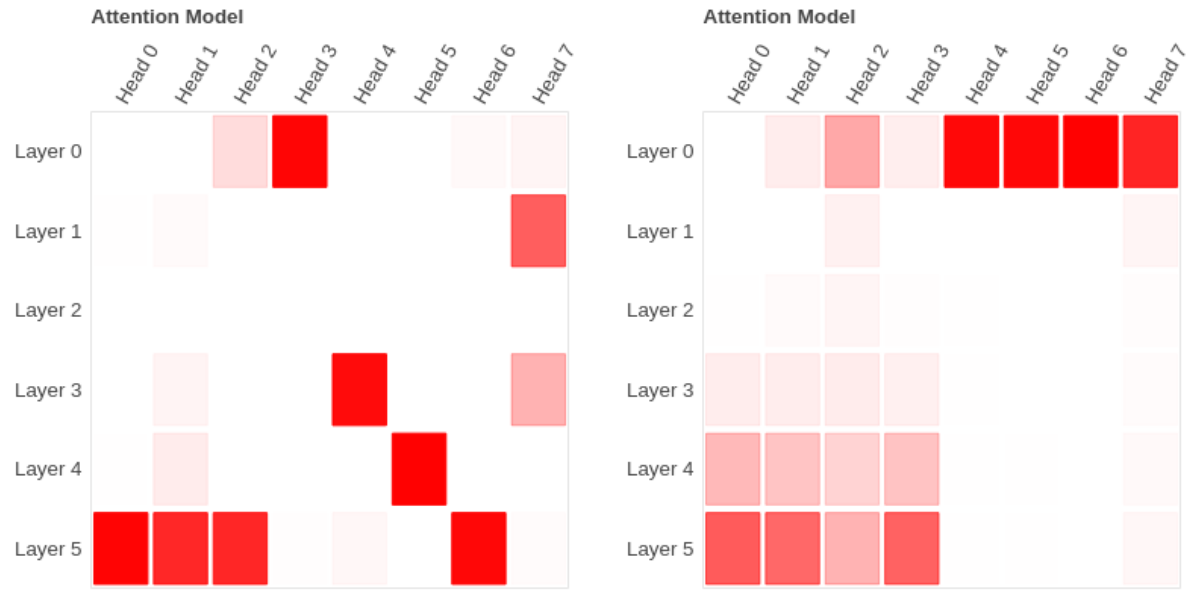

Figure 21: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder self-attention for the input sentence: "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

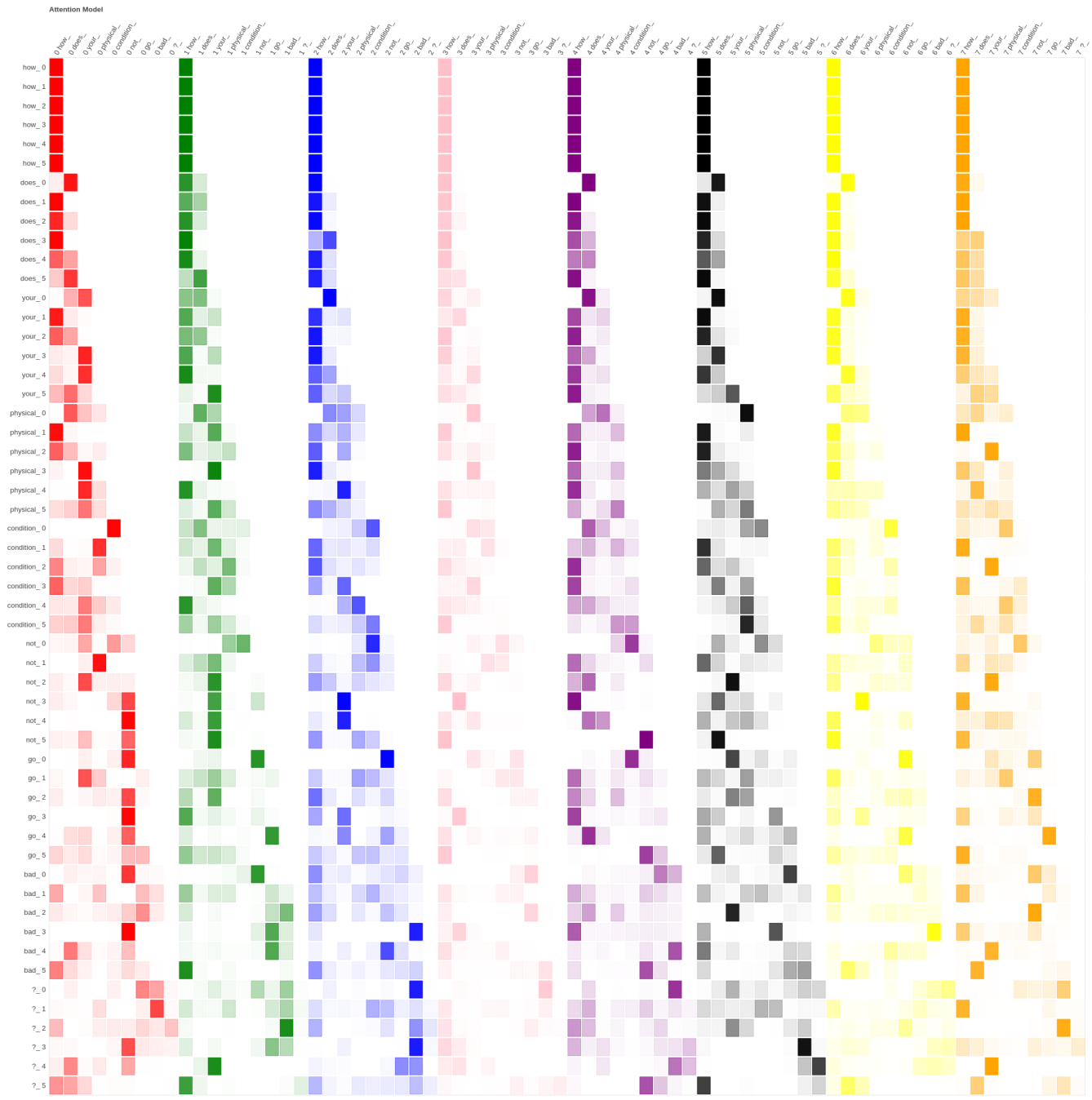

Figure 22: Vanilla 6-layer NMT model's attention heat-map for the decoder self-attention for the generated target sentence: "how does your physical condition not go bad ?". The input sentence is "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

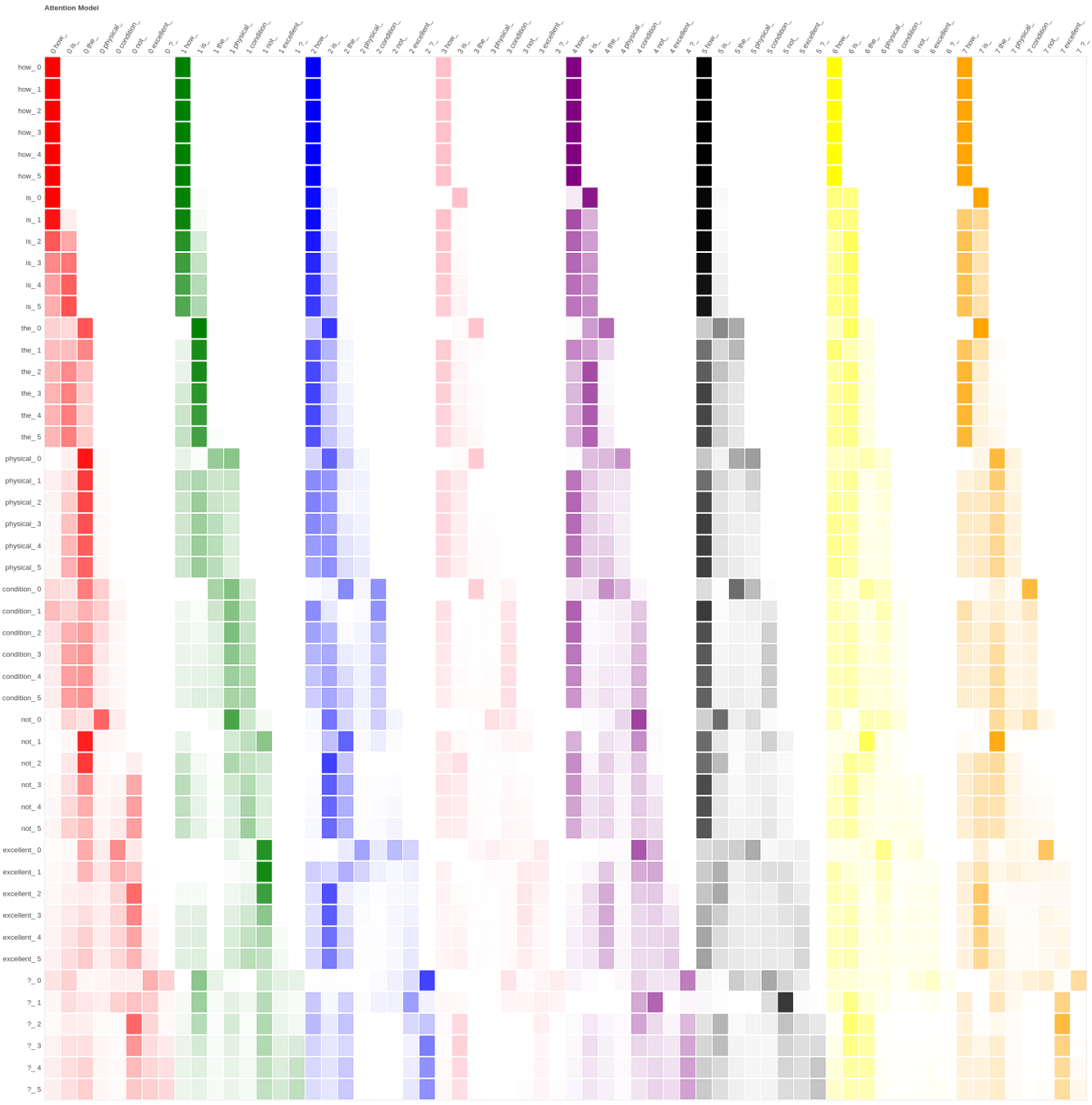

Figure 23: 6-layer RS-NMT model's attention heat-map for the encoder-decoder cross-attention for the generated target sentence: "how is the physical condition not excellent ?". The input sentence is "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

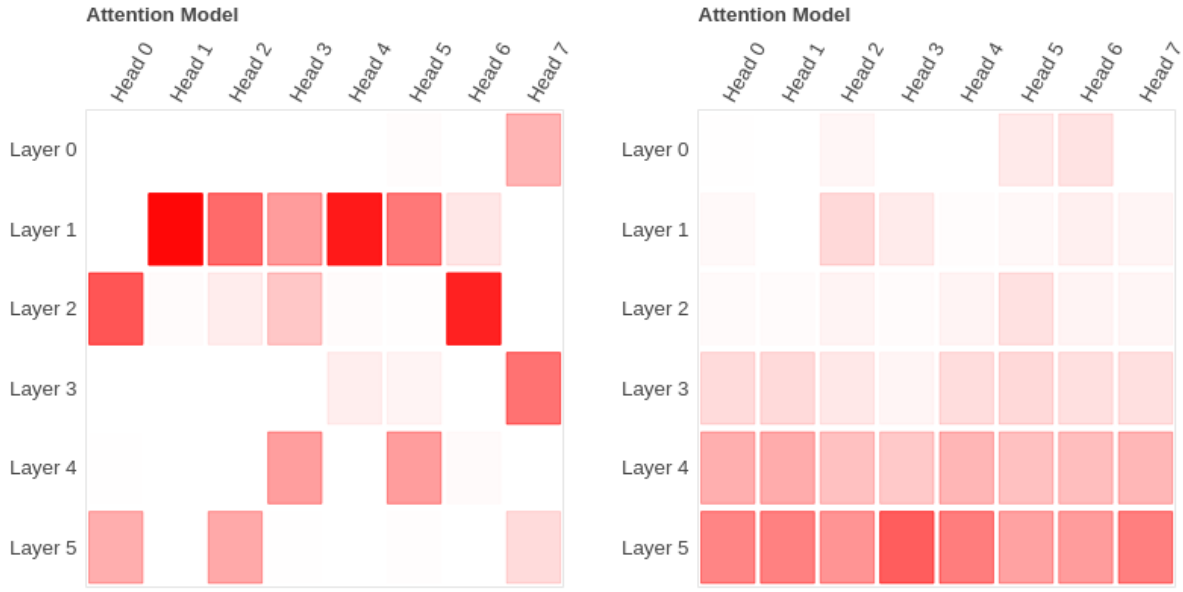

Figure 24: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) for the decoder self-attention for the generated target sentences: "how does your physical condition not go bad ?" and "how is the physical condition not excellent ?". The input sentence is: "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

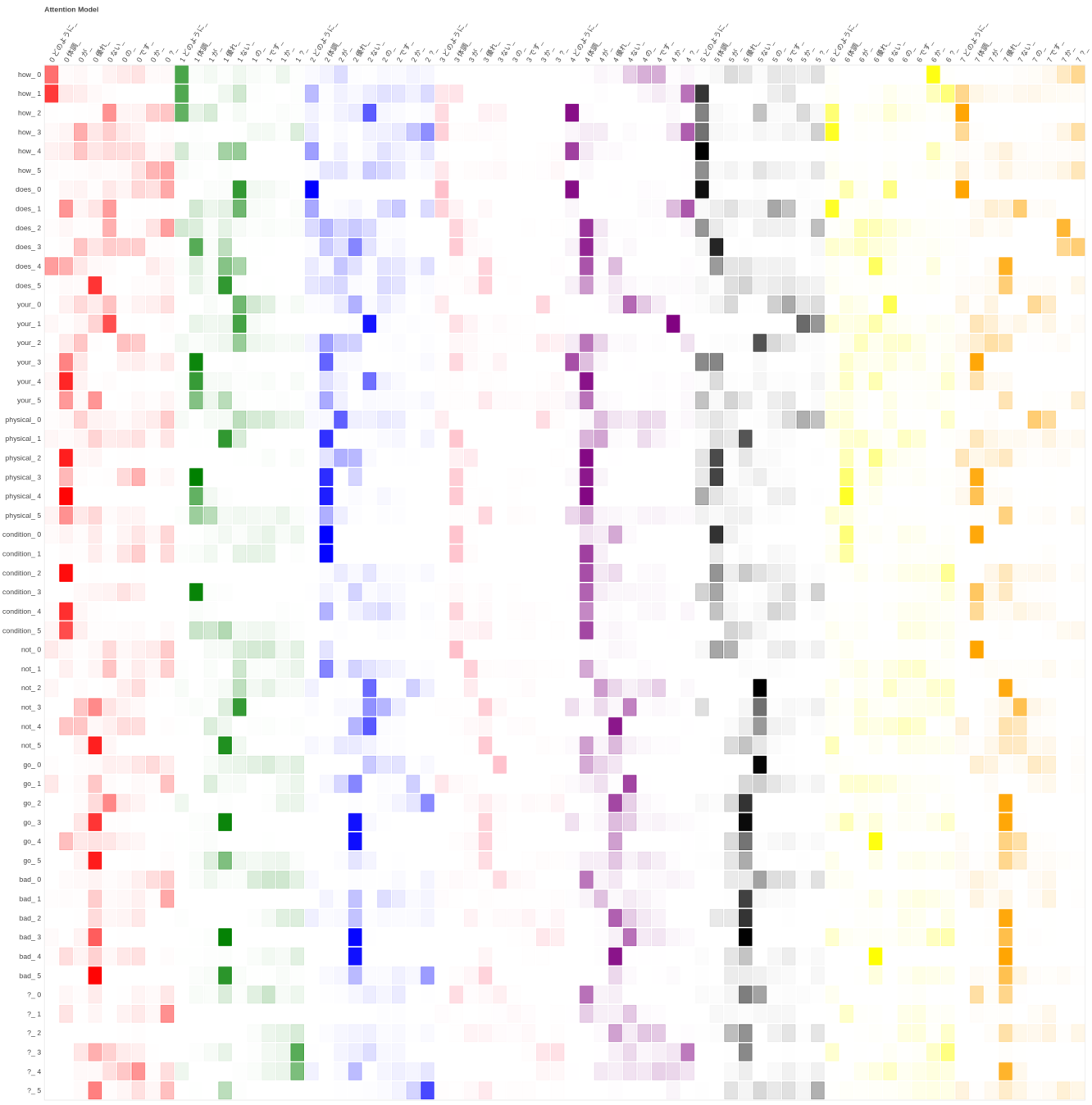

Figure 25: Vanilla 6-layer NMT model's attention heat-map for the encoder-decoder cross-attention for the generated target sentence: "how does your physical condition not go bad". The input sentence is "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

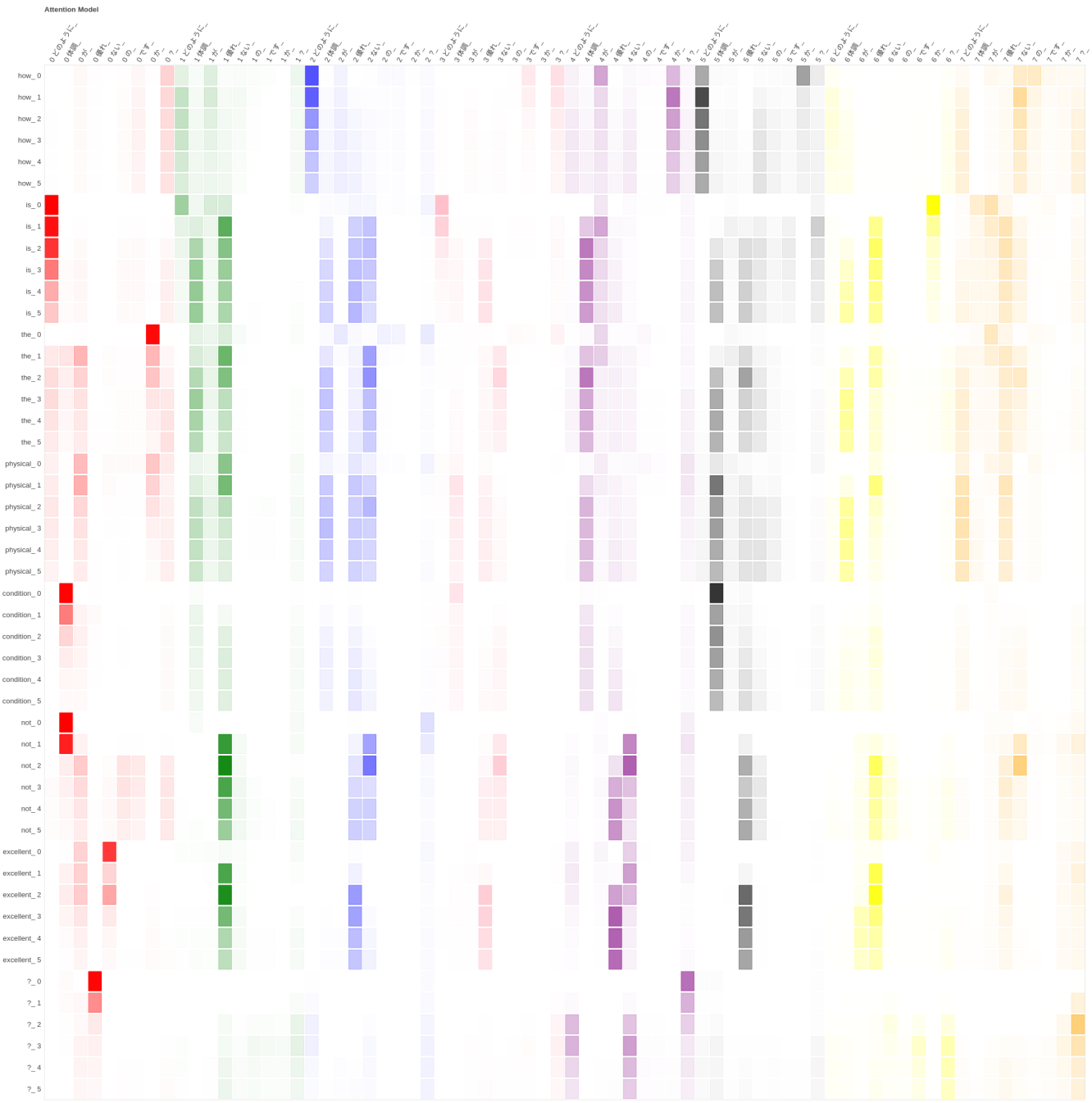

Figure 26: 6-layer RS-NMT model's attention heat-map for the encoder-decoder cross-attention for the generated target sentence: "how is the physical condition not excellent ?". The input sentence is "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).

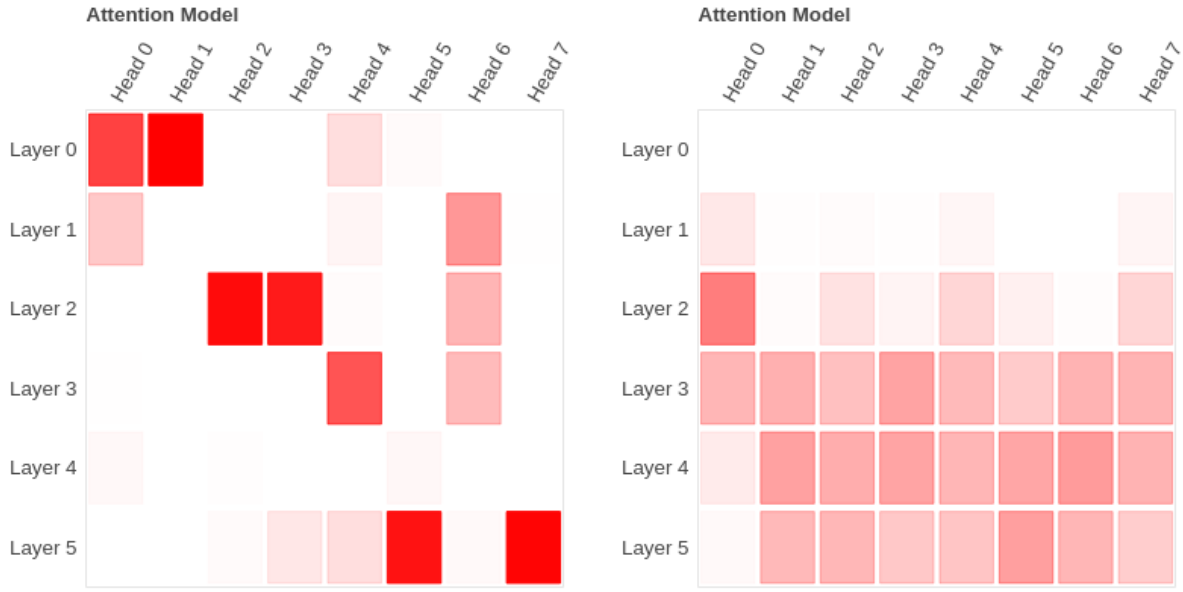

Figure 27: Entropy heat-maps for the vanilla 6-layer NMT model (left) and 6-layer RS-NMT model's (right) encoder-decoder cross-attention for the generated target sentences. "how does your physical condition not go bad ?" and "how is the physical condition not excellent ?". The input sentence is: "どのように 体調 が 優れ ない の です か ?" (How don't you feel well?).