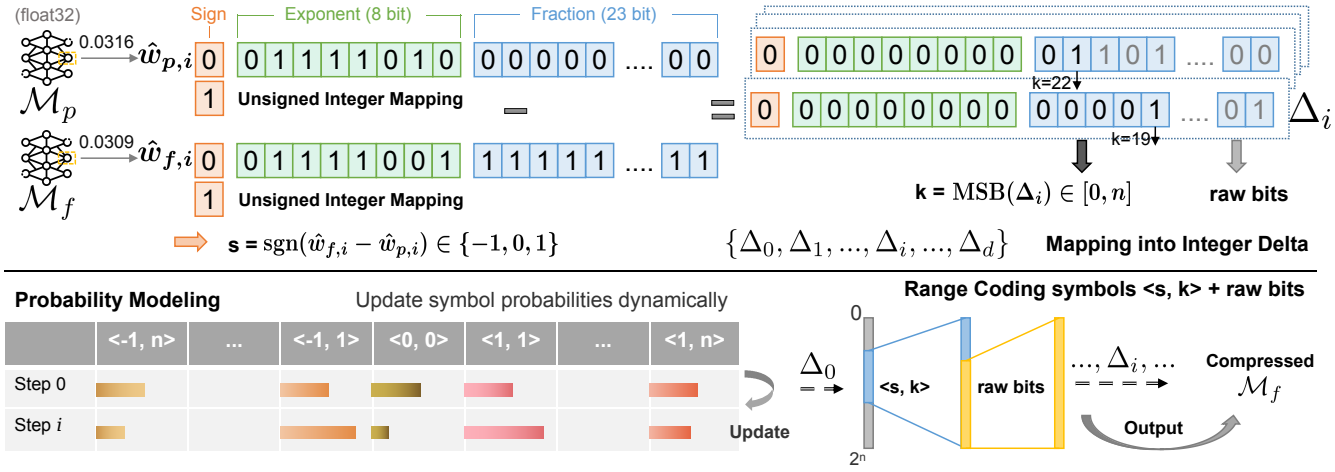

Figure 1: The lossless compression workflow of `FM-Delta`. The `FM-Delta` scheme (1) maps the two floating-point parameter elements at the same position of fine-tuned and pre-trained models into unsigned integers, and performs integer subtraction to obtain the bit-redundant delta element. Then it (2) regards the sign s and the most significant bit k of delta as symbols. With a quasi-static probability modeler, it encodes the symbols and scales the range to involve raw bits on all delta elements, leading to the compressed fine-tuned model.

Table 1: `FM-Delta`, int8 delta quantization and int4 delta quantization on two user-uploaded GPT2 models from Huggingface. "vicgalle/gpt2-alpaca-gpt" is a popular model with detailed meta information, and "jacksee/gpt2-finetuned-biochemistry" is an unpopular model without additional meta information. It should be mentioned that the embedding layer is not compressed due to shape inconsistency.

Model	Comp. Rate	MNLI (acc↑)	HellaSwag (acc↑)	ARC (acc↑)	WikiText2 (ppl↓)
gpt2	-	33.86	28.91	43.86	37.3835
vicgalle/gpt2-alpaca-gpt4 (lossless)	65.4% (ours)	33.90	29.16	45.03	44.7223
vicgalle/gpt2-alpaca-gpt4 (int8 delta)	65.5%	33.88	29.15	45.08	44.7078
vicgalle/gpt2-alpaca-gpt4 (int4 delta)	48.2%	34.11	29.18	45.03	45.3403
jacksee/gpt2-finetuned-biochemistry (lossless)	59.4% (ours)	31.95	27.46	40.40	24564191.2047
jacksee/gpt2-finetuned-biochemistry (int8 delta)	65.5%	32.01	27.46	40.45	25430731.7616
jacksee/gpt2-finetuned-biochemistry (int4 delta)	48.2%	32.85	27.23	38.51	1355776245.9481