( \*\*kwargs )

**Parameters that control the length of the output**

* **max\_length** (int, *optional*, defaults to 20) — The maximum length the generated tokens can have. Corresponds to the length of the input prompt + max\_new\_tokens. Its effect is overridden by max\_new\_tokens, if also set.
* **max\_new\_tokens** (int, *optional*) — The maximum numbers of tokens to generate, ignoring the number of tokens in the prompt.
* **min\_length** (int, *optional*, defaults to 0) — The minimum length of the sequence to be generated. Corresponds to the length of the input prompt + min\_new\_tokens. Its effect is overridden by min\_new\_tokens, if also set.
* **min\_new\_tokens** (int, *optional*) — The minimum numbers of tokens to generate, ignoring the number of tokens in the prompt.
* **early\_stopping** (bool or str, *optional*, defaults to False) — Controls the stopping condition for beam-based methods, like beam-search. It accepts the following values: True, where the generation stops as soon as there are num\_beams complete candidates; False, where an heuristic is applied and the generation stops when is it very unlikely to find better candidates; "never", where the beam search procedure only stops when there cannot be better candidates (canonical beam search algorithm).
* **max\_time** (float, *optional*) — The maximum amount of time you allow the computation to run for in seconds. generation will still finish the current pass after allocated time has been passed.
* **stop\_strings** (str or List[str], *optional*) — A string or a list of strings that should terminate generation if the model outputs them.

**Parameters that control the generation strategy used**

* **do\_sample** (bool, *optional*, defaults to False) — Whether or not to use sampling ; use greedy decoding otherwise.
* **num\_beams** (int, *optional*, defaults to 1) — Number of beams for beam search. 1 means no beam search.
* **num\_beam\_groups** (int, *optional*, defaults to 1) — Number of groups to divide num\_beams into in order to ensure diversity among different groups of beams. [this paper](https://arxiv.org/pdf/1610.02424.pdf) for more details.
* **penalty\_alpha** (float, *optional*) — The values balance the model confidence and the degeneration penalty in contrastive search decoding.
* **dola\_layers** (str or List[int], *optional*) — The layers to use for DoLa decoding. If None, DoLa decoding is not used. If a string, it must be one of “low” or “high”, which means using the lower part or higher part of the model layers, respectively. “low” means the first half of the layers up to the first 20 layers, and “high” means the last half of the layers up to the last 20 layers. If a list of integers, it must contain the indices of the layers to use for candidate premature layers in DoLa. The 0-th layer is the word embedding layer of the model. Set to 'low' to improve long-answer reasoning tasks, 'high' to improve short-answer tasks. Check the [documentation](https://github.com/huggingface/transformers/blob/main/docs/source/en/generation_strategies.md) or [the paper](https://arxiv.org/abs/2309.03883) for more details.

**Parameters that control the cache**

* **use\_cache** (bool, *optional*, defaults to True) — Whether or not the model should use the past last key/values attentions (if applicable to the model) to speed up decoding.
* **cache\_implementation** (str, *optional*, default to None) — Name of the cache class that will be instantiated in generate, for faster decoding. Possible values are:
  + "static": [StaticCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.StaticCache)
  + "offloaded\_static": [OffloadedStaticCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.OffloadedStaticCache)
  + "sliding\_window": [SlidingWindowCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.SlidingWindowCache)
  + "hybrid": [HybridCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.HybridCache)
  + "mamba": [MambaCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.MambaCache)
  + "quantized": [QuantizedCache](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.QuantizedCache)

We support other cache types, but they must be manually instantiated and passed to generate through the past\_key\_values argument. See our [cache documentation](https://huggingface.co/docs/transformers/en/kv_cache) for further information.

* **cache\_config** (CacheConfig or dict, *optional*, default to None) — Arguments used in the key-value cache class can be passed in cache\_config. Can be passed as a Dict and it will be converted to its repsective CacheConfig internally. Otherwise can be passed as a CacheConfig class matching the indicated cache\_implementation.
* **return\_legacy\_cache** (bool, *optional*, default to True) — Whether to return the legacy or new format of the cache when DynamicCache is used by default.

**Parameters for manipulation of the model output logits**

* **temperature** (float, *optional*, defaults to 1.0) — The value used to modulate the next token probabilities.
* **top\_k** (int, *optional*, defaults to 50) — The number of highest probability vocabulary tokens to keep for top-k-filtering.
* **top\_p** (float, *optional*, defaults to 1.0) — If set to float < 1, only the smallest set of most probable tokens with probabilities that add up to top\_p or higher are kept for generation.
* **min\_p** (float, *optional*) — Minimum token probability, which will be scaled by the probability of the most likely token. It must be a value between 0 and 1. Typical values are in the 0.01-0.2 range, comparably selective as setting top\_p in the 0.99-0.8 range (use the opposite of normal top\_p values).
* **typical\_p** (float, *optional*, defaults to 1.0) — Local typicality measures how similar the conditional probability of predicting a target token next is to the expected conditional probability of predicting a random token next, given the partial text already generated. If set to float < 1, the smallest set of the most locally typical tokens with probabilities that add up to typical\_p or higher are kept for generation. See [this paper](https://arxiv.org/pdf/2202.00666.pdf) for more details.
* **epsilon\_cutoff** (float, *optional*, defaults to 0.0) — If set to float strictly between 0 and 1, only tokens with a conditional probability greater than epsilon\_cutoff will be sampled. In the paper, suggested values range from 3e-4 to 9e-4, depending on the size of the model. See [Truncation Sampling as Language Model Desmoothing](https://arxiv.org/abs/2210.15191) for more details.
* **eta\_cutoff** (float, *optional*, defaults to 0.0) — Eta sampling is a hybrid of locally typical sampling and epsilon sampling. If set to float strictly between 0 and 1, a token is only considered if it is greater than either eta\_cutoff or sqrt(eta\_cutoff) \* exp(-entropy(softmax(next\_token\_logits))). The latter term is intuitively the expected next token probability, scaled by sqrt(eta\_cutoff). In the paper, suggested values range from 3e-4 to 2e-3, depending on the size of the model. See [Truncation Sampling as Language Model Desmoothing](https://arxiv.org/abs/2210.15191) for more details.
* **diversity\_penalty** (float, *optional*, defaults to 0.0) — This value is subtracted from a beam’s score if it generates a token same as any beam from other group at a particular time. Note that diversity\_penalty is only effective if group beam search is enabled.
* **repetition\_penalty** (float, *optional*, defaults to 1.0) — The parameter for repetition penalty. 1.0 means no penalty. See [this paper](https://arxiv.org/pdf/1909.05858.pdf) for more details.
* **encoder\_repetition\_penalty** (float, *optional*, defaults to 1.0) — The paramater for encoder\_repetition\_penalty. An exponential penalty on sequences that are not in the original input. 1.0 means no penalty.
* **length\_penalty** (float, *optional*, defaults to 1.0) — Exponential penalty to the length that is used with beam-based generation. It is applied as an exponent to the sequence length, which in turn is used to divide the score of the sequence. Since the score is the log likelihood of the sequence (i.e. negative), length\_penalty > 0.0 promotes longer sequences, while length\_penalty < 0.0 encourages shorter sequences.
* **no\_repeat\_ngram\_size** (int, *optional*, defaults to 0) — If set to int > 0, all ngrams of that size can only occur once.
* **bad\_words\_ids** (List[List[int]], *optional*) — List of list of token ids that are not allowed to be generated. Check [NoBadWordsLogitsProcessor](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.NoBadWordsLogitsProcessor) for further documentation and examples.
* **force\_words\_ids** (List[List[int]] or List[List[List[int]]], *optional*) — List of token ids that must be generated. If given a List[List[int]], this is treated as a simple list of words that must be included, the opposite to bad\_words\_ids. If given List[List[List[int]]], this triggers a [disjunctive constraint](https://github.com/huggingface/transformers/issues/14081), where one can allow different forms of each word.
* **renormalize\_logits** (bool, *optional*, defaults to False) — Whether to renormalize the logits after applying all the logits processors (including the custom ones). It’s highly recommended to set this flag to True as the search algorithms suppose the score logits are normalized but some logit processors break the normalization.
* **constraints** (List[Constraint], *optional*) — Custom constraints that can be added to the generation to ensure that the output will contain the use of certain tokens as defined by Constraint objects, in the most sensible way possible.
* **forced\_bos\_token\_id** (int, *optional*, defaults to model.config.forced\_bos\_token\_id) — The id of the token to force as the first generated token after the decoder\_start\_token\_id. Useful for multilingual models like [mBART](https://huggingface.co/docs/transformers/model_doc/mbart) where the first generated token needs to be the target language token.
* **forced\_eos\_token\_id** (int or List[int], \*optional\*, defaults to model.config.forced\_eos\_token\_id) -- The id of the token to force as the last generated token when max\_length` is reached. Optionally, use a list to set multiple *end-of-sequence* tokens.
* **remove\_invalid\_values** (bool, *optional*, defaults to model.config.remove\_invalid\_values) — Whether to remove possible *nan* and *inf* outputs of the model to prevent the generation method to crash. Note that using remove\_invalid\_values can slow down generation.
* **exponential\_decay\_length\_penalty** (tuple(int, float), *optional*) — This Tuple adds an exponentially increasing length penalty, after a certain amount of tokens have been generated. The tuple shall consist of: (start\_index, decay\_factor) where start\_index indicates where penalty starts and decay\_factor represents the factor of exponential decay
* **suppress\_tokens** (List[int], *optional*) — A list of tokens that will be suppressed at generation. The SupressTokens logit processor will set their log probs to -inf so that they are not sampled.
* **begin\_suppress\_tokens** (List[int], *optional*) — A list of tokens that will be suppressed at the beginning of the generation. The SupressBeginTokens logit processor will set their log probs to -inf so that they are not sampled.
* **forced\_decoder\_ids** (List[List[int]], *optional*) — A list of pairs of integers which indicates a mapping from generation indices to token indices that will be forced before sampling. For example, [[1, 123]] means the second generated token will always be a token of index 123.
* **sequence\_bias** (Dict[Tuple[int], float], *optional*)) — Dictionary that maps a sequence of tokens to its bias term. Positive biases increase the odds of the sequence being selected, while negative biases do the opposite. Check [SequenceBiasLogitsProcessor](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.SequenceBiasLogitsProcessor) for further documentation and examples.
* **token\_healing** (bool, *optional*, defaults to False) — Heal tail tokens of prompts by replacing them with their appropriate extensions. This enhances the quality of completions for prompts affected by greedy tokenization bias.
* **guidance\_scale** (float, *optional*) — The guidance scale for classifier free guidance (CFG). CFG is enabled by setting guidance\_scale > 1. Higher guidance scale encourages the model to generate samples that are more closely linked to the input prompt, usually at the expense of poorer quality.
* **low\_memory** (bool, *optional*) — Switch to sequential beam search and sequential topk for contrastive search to reduce peak memory. Used with beam search and contrastive search.
* **watermarking\_config** (BaseWatermarkingConfig or dict, *optional*) — Arguments used to watermark the model outputs by adding a small bias to randomly selected set of “green” tokens. See the docs of [SynthIDTextWatermarkingConfig](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.SynthIDTextWatermarkingConfig) and [WatermarkingConfig](https://huggingface.co/docs/transformers/v4.47.1/en/internal/generation_utils" \l "transformers.WatermarkingConfig) for more details. If passed as Dict, it will be converted to a WatermarkingConfig internally.

**Parameters that define the output variables of generate**

* **num\_return\_sequences** (int, *optional*, defaults to 1) — The number of independently computed returned sequences for each element in the batch.
* **output\_attentions** (bool, *optional*, defaults to False) — Whether or not to return the attentions tensors of all attention layers. See attentions under returned tensors for more details.
* **output\_hidden\_states** (bool, *optional*, defaults to False) — Whether or not to return the hidden states of all layers. See hidden\_states under returned tensors for more details.
* **output\_scores** (bool, *optional*, defaults to False) — Whether or not to return the prediction scores. See scores under returned tensors for more details.
* **output\_logits** (bool, *optional*) — Whether or not to return the unprocessed prediction logit scores. See logits under returned tensors for more details.
* **return\_dict\_in\_generate** (bool, *optional*, defaults to False) — Whether or not to return a [ModelOutput](https://huggingface.co/docs/transformers/v4.47.1/en/main_classes/output" \l "transformers.utils.ModelOutput), as opposed to returning exclusively the generated sequence. This flag must be set to True to return the generation cache (when use\_cache is True) or optional outputs (see flags starting with output\_)

**Special tokens that can be used at generation time**

* **pad\_token\_id** (int, *optional*) — The id of the *padding* token.
* **bos\_token\_id** (int, *optional*) — The id of the *beginning-of-sequence* token.
* **eos\_token\_id** (Union[int, List[int]], *optional*) — The id of the *end-of-sequence* token. Optionally, use a list to set multiple *end-of-sequence* tokens.

**Generation parameters exclusive to encoder-decoder models**

* **encoder\_no\_repeat\_ngram\_size** (int, *optional*, defaults to 0) — If set to int > 0, all ngrams of that size that occur in the encoder\_input\_ids cannot occur in the decoder\_input\_ids.
* **decoder\_start\_token\_id** (int or List[int], *optional*) — If an encoder-decoder model starts decoding with a different token than *bos*, the id of that token or a list of length batch\_size. Indicating a list enables different start ids for each element in the batch (e.g. multilingual models with different target languages in one batch)

**Generation parameters exclusive to assistant generation**

* **is\_assistant** (bool, *optional*, defaults to False) — Whether the model is an assistant (draft) model.
* **num\_assistant\_tokens** (int, *optional*, defaults to 20) — Defines the number of *speculative tokens* that shall be generated by the assistant model before being checked by the target model at each iteration. Higher values for num\_assistant\_tokens make the generation more *speculative* : If the assistant model is performant larger speed-ups can be reached, if the assistant model requires lots of corrections, lower speed-ups are reached.
* **num\_assistant\_tokens\_schedule** (str, *optional*, defaults to "constant") — Defines the schedule at which max assistant tokens shall be changed during inference.
  + "heuristic": When all speculative tokens are correct, increase num\_assistant\_tokens by 2 else reduce by 1. num\_assistant\_tokens value is persistent over multiple generation calls with the same assistant model.
  + "heuristic\_transient": Same as "heuristic" but num\_assistant\_tokens is reset to its initial value after each generation call.
  + "constant": num\_assistant\_tokens stays unchanged during generation
* **assistant\_confidence\_threshold** (float, *optional*, defaults to 0.4) — The confidence threshold for the assistant model. If the assistant model’s confidence in its prediction for the current token is lower than this threshold, the assistant model stops the current token generation iteration, even if the number of *speculative tokens* (defined by num\_assistant\_tokens) is not yet reached. The assistant’s confidence threshold is adjusted throughout the speculative iterations to reduce the number of unnecessary draft and target forward passes, biased towards avoiding false negatives. assistant\_confidence\_threshold value is persistent over multiple generation calls with the same assistant model. It is an unsupervised version of the dynamic speculation lookahead from Dynamic Speculation Lookahead Accelerates Speculative Decoding of Large Language Models <https://arxiv.org/abs/2405.04304>.
* **prompt\_lookup\_num\_tokens** (int, *optional*) — The number of tokens to be output as candidate tokens.
* **max\_matching\_ngram\_size** (int, *optional*) — The maximum ngram size to be considered for matching in the prompt. Default to 2 if not provided.
* **assistant\_early\_exit(int,** *optional*) — If set to a positive integer, early exit of the model will be used as an assistant. Can only be used with models that support early exit (i.e. models where logits from intermediate layers can be interpreted by the LM head).
* **assistant\_lookbehind(int,** *optional*, defaults to 10) — If set to a positive integer, the re-encodeing process will additionally consider the last assistant\_lookbehind assistant tokens to correctly align tokens. Can only be used with different tokenizers in speculative decoding. See this [blog](https://huggingface.co/blog/universal_assisted_generation) for more details.
* **target\_lookbehind(int,** *optional*, defaults to 10) — If set to a positive integer, the re-encodeing process will additionally consider the last target\_lookbehind target tokens to correctly align tokens. Can only be used with different tokenizers in speculative decoding. See this [blog](https://huggingface.co/blog/universal_assisted_generation) for more details.

**Parameters related to performances and compilation**

* **compile\_config** (CompileConfig, *optional*) — If using a static cache, this controls how generate will compile the forward pass for performance gains.

**Wild card**

* **generation\_kwargs** — Additional generation kwargs will be forwarded to the generate function of the model. Kwargs that are not present in generate’s signature will be used in the model forward pass.

Class that holds a configuration for a generation task. A generate call supports the following generation methods for text-decoder, text-to-text, speech-to-text, and vision-to-text models:

* *greedy decoding* if num\_beams=1 and do\_sample=False
* *contrastive search* if penalty\_alpha>0. and top\_k>1
* *multinomial sampling* if num\_beams=1 and do\_sample=True
* *beam-search decoding* if num\_beams>1 and do\_sample=False
* *beam-search multinomial sampling* if num\_beams>1 and do\_sample=True
* *diverse beam-search decoding* if num\_beams>1 and num\_beam\_groups>1
* *constrained beam-search decoding* if constraints!=None or force\_words\_ids!=None
* *assisted decoding* if assistant\_model or prompt\_lookup\_num\_tokens is passed to .generate()
* *dola decoding* if dola\_layers is passed to .generate()