|  |  |  |  |
| --- | --- | --- | --- |
| Task | Template type | Prompt template | Expected response |
| VQA | Extraction | "document} {quest ion}" | answer annotation |
| NLI | MCQ | " document } V{statement }", Yes or No?" | answer annotation |
| KIE  MCQ  Internal classification | Extraction | "{ document } What is the value for the I"{key}"?" | Associated value annotation |
| "{document } What is I"{value}\" in the document? Possible choices: { choices}." (where choices is a subset of allthe keys in the dataset in random order). | Associated key annotation |  |
| "{document } What is I"{value}" in the document?"? | Associated key annotation |  |
| CLS  Internal classification | MCQ | "{document } What type of document is this? Possible choices: {choices}." (where choices is a subset of all the classes in the dataset in random order). | class annotation |
| "document } What type of document is this?" | class annotation |  |

contains a consecutive series of tokens. Further, let x be a corrupted version of x where the contiguous tokens corresponding to a sampled text block are replaced with a special mask token [M]. To facilitate the identification of the block to be filled during text generation, each input block is augmented with a special start token [S] while the. output block includes an end token [E]. For instance, a block with tokens (x4, x5) becomes [M] in x, ([S], x4, x5. when conditioned upon, and is expected to generate (x4, x5, E) as output autoregressively (see Figure|2[for a detailed illustration of these configurations). The following cross-entropy loss is then minimized for the infilling objective..

# 3.4Instruction Tuning

Following recent work in the field of VRDU [12,[31, 32] and prior work in NLP [40] 41], we instruction-tune DocLLM on a variety of instructions derived from DocAI datasets using various templates. Due to the high cost and time intensity of manual data collection, we leave the construction of a VRDU instruction-tuning dataset with crowdsourced instructions and preferences to future work. We employ a total of 16 datasets with their corresponding OCRs, spanning four DocAI tasks: visual question answering (VQA), natural language inference (NLI), key information extraction (KIE), and document classification (CLS)

The diversity of supervised fine tuning (SFT) instructions is critical in helping zero-shot generalization [40, 41, 2] Thus, we diversify templates per task when possible, with each template asking a different question, and in some cases. expecting different types of answers. We re-use the templates introduced in [31] 32] when applicable, and consider a broader selection of datasets in our instruction-tuning data mix..

We create the templates following what we believe end users would generally ask about documents (Table|1b. For KIE and CLS, we hypothesize that (1) the extraction instructions can teach DocLLM to correlate names of keys in the prompts. with document fields so as to retrieve values, (2) the internal classification instructions can help the model understand what intrinsically characterizes each key or document type, and (3) the multiple choice question (MCQ) instructions. can teach the model to leverage its comprehension of key names included as choices in the prompt (resp. document type names) to classify extracted values (resp. entire documents). We introduce the templates in detail as follows..