$c_t = a_t c_{t-1} + z_t$ から始める自作LLM

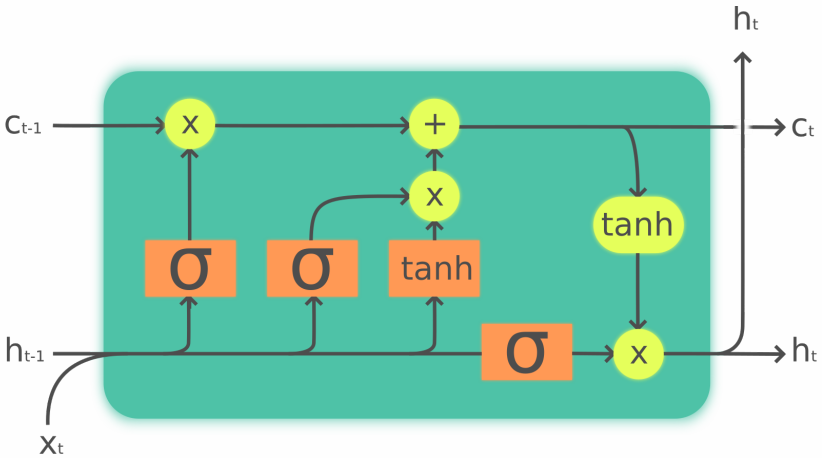
時系列モデルSioConvの紹介

myxy

2024-07-24

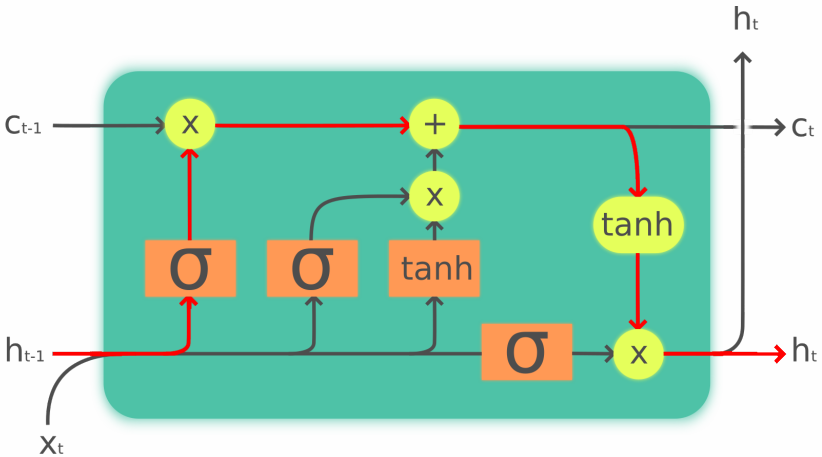
1. LSTM
2. LSTM の簡略化
3. 忘却ゲートの並列計算
4. SioConv
5. 実験（事前学習編）
6. 実験（会話応答ファインチューニング編）

LSTM



LSTM は Transformer 以前に時系列処理に用いられてきた

LSTM



"The LSTM Cell" (© Guillaume Chevalier, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>) を改変して作成

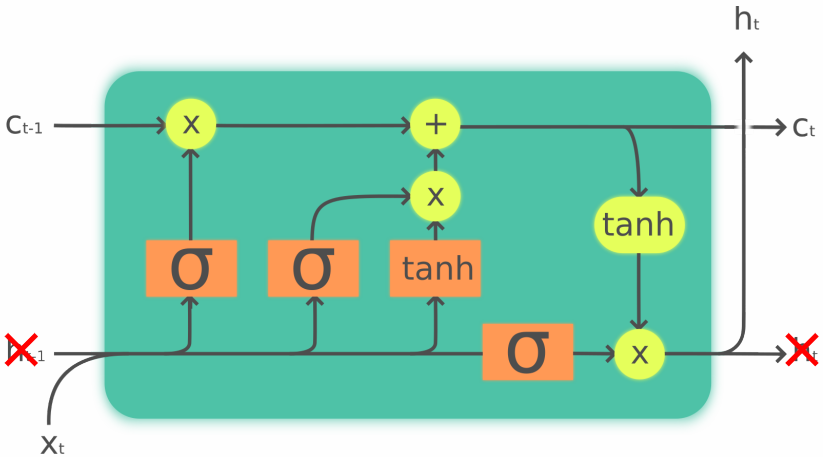
隠れ状態 h の伝搬経路に全結合層（図の σ）が存在

LSTMの課題

LSTMで1000ステップの時系列を予測する場合
全結合層を1000回計算しなければならない
Transformerのように時系列を並列に扱うことが困難

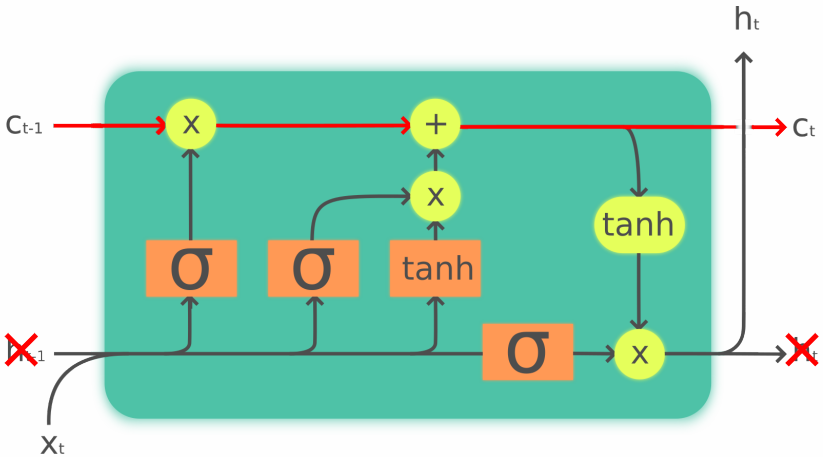
→LSTMを簡略化して並列化できるだろうか？

隠れ状態hの除去



伝搬経路上に全結合層がある隠れ状態hを除去する

隠れ状態 h の除去



隠れ状態 c の伝搬はスカラー毎の一次式となる

LSTM の忘却ゲートを $a_t = Sigmoid(Linear(x_t))$、忘却ゲートを
$z_t = Sigmoid(Linear(x_t) \bigodot Tanh(Linear(x_t))$ としたとき
隠れ状態 c_t は $c_t = a_t c_{t-1} + z_t$ の形に書くことができる

計算

例として長さ 4 の系列を計算してみる

$$
\begin{aligned}
c_0 &= a_0 c_{-1} + z_0 \\
c_1 &= a_1 a_0 c_{-1} + a_1 z_0 + z_1 \\
c_2 &= a_2 a_1 a_0 c_{-1} + a_2 a_1 z_0 + a_2 z_1 + z_2 \\
c_3 &= a_3 a_2 a_1 a_0 c_{-1} + a_3 a_2 a_1 z_0 + a_3 a_2 z_1 + a_3 z_2 + z_3
\end{aligned}
$$

行列の形式で書くと

$$
\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = c_{-1} \begin{bmatrix} a_0 \\ a_1 a_0 \\ a_2 a_1 a_0 \\ a_3 a_2 a_1 a_0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ a_1 & 1 & 0 & 0 \\ a_2 a_1 & a_2 & 1 & 0 \\ a_3 a_2 a_1 & a_3 a_2 & a_3 & 1 \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix}
$$

$$\begin{bmatrix} a_0 \\ a_1 a_0 \\ a_2 a_1 a_0 \\ a_3 a_2 a_1 a_0 \end{bmatrix}$$ や $$\begin{bmatrix} 1 & 0 & 0 & 0 \\ a_1 & 1 & 0 & 0 \\ a_2 a_1 & a_2 & 1 & 0 \\ a_3 a_2 a_1 & a_3 a_2 & a_3 & 1 \end{bmatrix}$$ をどのように計算するか

a_i の対数 $l_i = \ln a_i$ を考える

対数領域での累積和

$$\begin{bmatrix} a_0 \\ a_1 a_0 \\ a_2 a_1 a_0 \\ a_3 a_2 a_1 a_0 \end{bmatrix} = \exp \begin{bmatrix} l_0 \\ l_1 + l_0 \\ l_2 + l_1 + l_0 \\ l_3 + l_2 + l_1 + l_0 \end{bmatrix}$$

$\exp$ の中身は $\begin{bmatrix} l_0 \\ l_1 \\ l_2 \\ l_3 \end{bmatrix}$ の累積和 (`torch.cumsum()`) として書ける

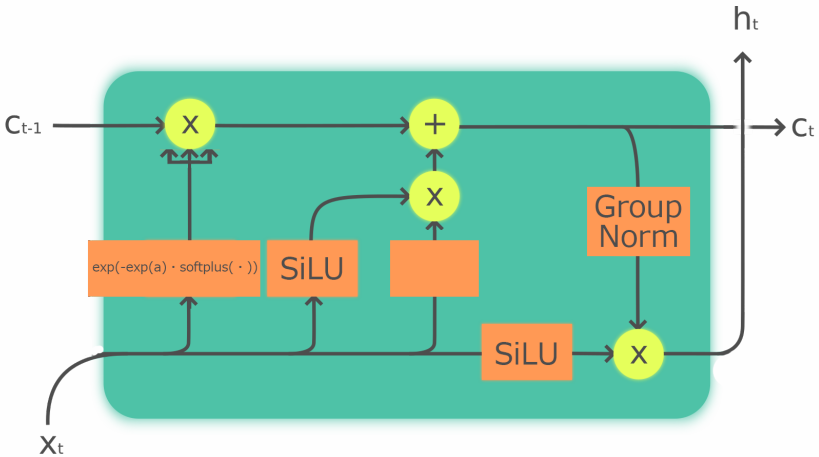
対数領域での累積和

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ a_1 & 1 & 0 & 0 \\ a_2 a_1 & a_2 & 1 & 0 \\ a_3 a_2 a_1 & a_3 a_2 & a_3 & 1 \end{bmatrix} = \exp \begin{bmatrix} 0 & -\infty & -\infty & -\infty \\ l_1 & 0 & -\infty & -\infty \\ l_2 + l_1 & l_2 & 0 & -\infty \\ l_3 + l_2 + l_1 & l_3 + l_2 & l_3 & 0 \end{bmatrix}$$

exp の中身の下三角部分は $\begin{bmatrix} 0 & 0 & 0 & 0 \\ l_1 & 0 & 0 & 0 \\ l_2 & l_2 & 0 & 0 \\ l_3 & l_3 & l_3 & 0 \end{bmatrix}$ の行方向の累積和で

この行列は `torch.tril()` を用いて作れる

SioConv



"The LSTM Cell" (© Guillaume Chevalier, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>) を改変して作成

忘却ゲートの並列化といくつかの変更を加えた

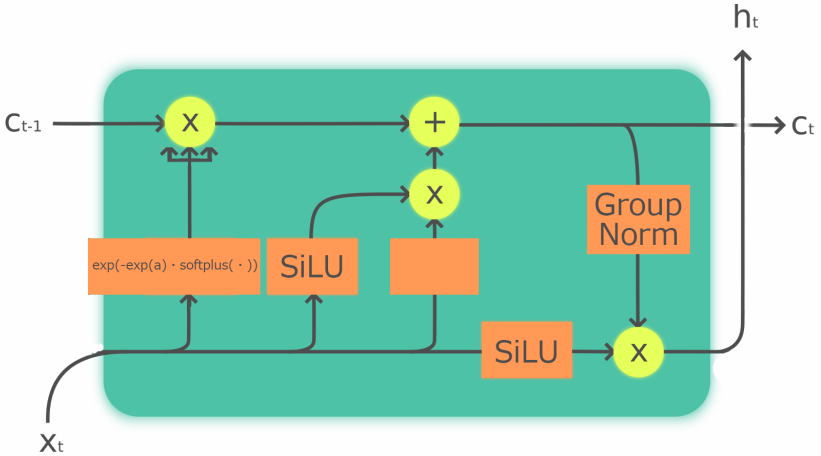
マルチヘッド化

忘却ゲートをチャンネル毎のスカラーにすると系列長 × 系列長のサイズの

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ a_1 & 1 & 0 & 0 & \cdots \\ a_2 a_1 & a_2 & 1 & 0 & \cdots \\ a_3 a_2 a_1 & a_3 a_2 & a_3 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

行列がチャンネル数分必要、流石にメモリが足りない
→ チャンネルをいくつかのヘッドに分割し、忘却ゲートはヘッド毎のスカラーとする

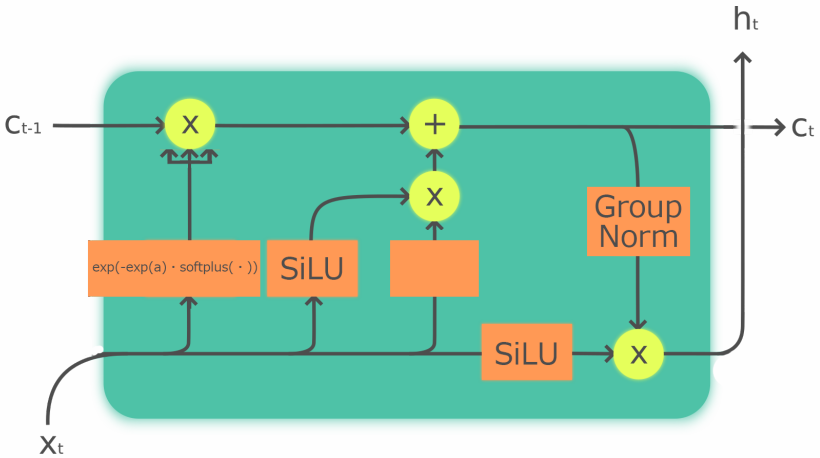
Gated Linear Unit



"The LSTM Cell" (© Guillaume Chevalier, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>) を改変して作成

LSTM では $tanh(Linear(x)) \cdot sigmoid(Linear(x))$ を採用していたが
GLU のアイデアに基づき $Linear(x) \cdot SiLU(Linear(x))$ のようにした

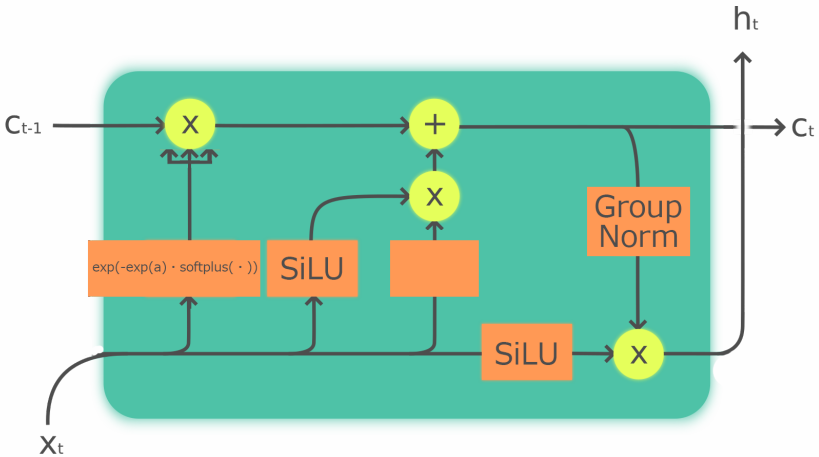
GroupNorm



"The LSTM Cell" (© Guillaume Chevalier, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>) を改変して作成

RetNet で用いられたヘッド毎の GroupNorm の採用

忘却ゲート



"The LSTM Cell" (© Guillaume Chevalier, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>) を改変して作成

忘却ゲートの計算は Mamba2 のソースコードを参考にした

実験

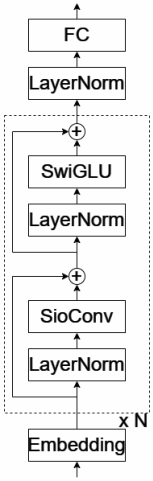
実際に SioConv を言語モデルとして学習させてみて
損失の下がり方や生成の様子を見てみる

学習モデル

Transformer と
類似したアーキテクチャを採用

- 次元数 2048
- ヘッド数 8
- ブロック数 36
- SwiGLU の隠れ次元数 4096

の設定で実験
約 1.6B パラメータ



学習データセット

- MiniPile (https://huggingface.co/datasets/JeanKaddour/minipile)
- 日本語 Wikipedia (https://huggingface.co/datasets/graelo/wikipedia)

を混合して使用
日英の自然言語文字列がいっぱい入ってる

トークナイザは ELYZA-japanese-Llama-2-7b-fast-instruct のものを使用（https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b-fast-instruct）

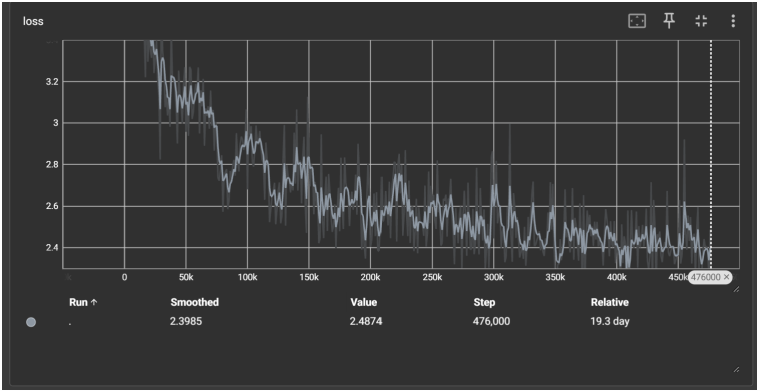
学習機材

- CPU:
  Ryzen Threadripper 1950X
- RAM: 128GB
- GPU: RTX3090 x5
  - VRAM 合計 120GB
  - 学習途中で 4->5 枚に増えた

その他学習設定

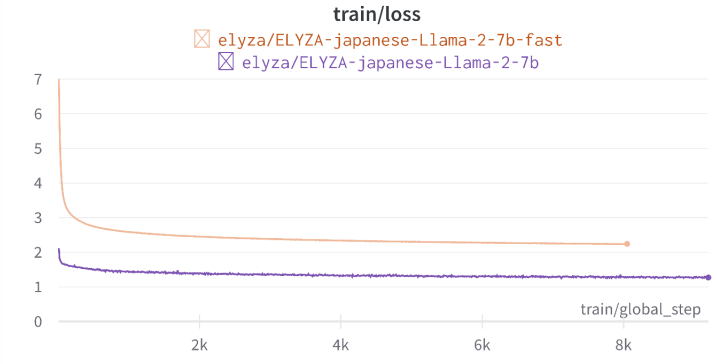
- オプティマイザ: AdamW
  - 学習率は 3e-4 から 1e-3 の範囲で 1000 ステップ周期で線形に上下
- コンテキスト長: 1024
  - 1024 トークン以上の文字列は切り捨て
- バッチサイズ: 50
  - 10 回に分けて勾配蓄積

学習曲線



約 19 日かけて学習（途中で電子レンジとか使うため止めたりはした）
Loss は 2.4 付近まで減少

既存の学習曲線との比較



https://zenn.dev/elyza/articles/2fd451c944649d より引用

同じトークナイザでの大規模学習では Loss は 2.2 程度まで減少

生成例（temperature=0.8）

prompt: 宮城県（
<s> 宮城県（みやぎけん）は、1948 年から 1963 年まで日本の県。現在の宮城県仙台市の一部にあたる。
本項では前身の宮城県立病院（みやぎけんびょういん）についても取り扱う。
概要
仙台県立宮城病院と仙台市立病院の統合
1948 年（昭和 23 年）の県立宮城病院開院に伴い、明治維新の功績を挙げた仙台市は、1949 年（昭和 24 年）に測量をメインとした府立仙台病院を開院した。しかし、この県立病院跡地は仙台市内の中心部から見える仙台市の中心部から見て南東方向に位置しており、標高は約 38 度で、東側は仙台底の海岸から急傾斜である。
1951 年（昭和 26 年）10 月、仙台市は国立病院機構整備法（昭和 26 年法律第 120 号）第 2 条第 1 項に基づき、仙台市立病院の指定を以て仙台市立病院を指定することとなり、仙台市・仙台市による同院病院医院の設置が確定した。同年 5 月、政府は仙台市と仙台市の中核都市である仙台市が、北都市と仙台市の中間のような立地であることや、国立仙台病院の移転を地方都市に都道府県庁が新設することを懸念して反発し、仙台市公立病院の設置も実現できず、このような状況に陥った。

生成の特徴

- Wikipedia をデータセットにしているので"宮城県（"のようなプロンプトを入れるとそれっぽい記事を生成してくれる
- 実際の情報との整合性はそんなにない

データセット

事前学習したモデルをさらに会話応答データセットでファインチューニングする

- https://huggingface.co/datasets/shi3z/ja_conv_wikipedia_llama2pro8b_30k
- https://huggingface.co/datasets/shi3z/ja_conv_wikipedia_orion14B_100K
- https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k

を混合して学習

学習の形式

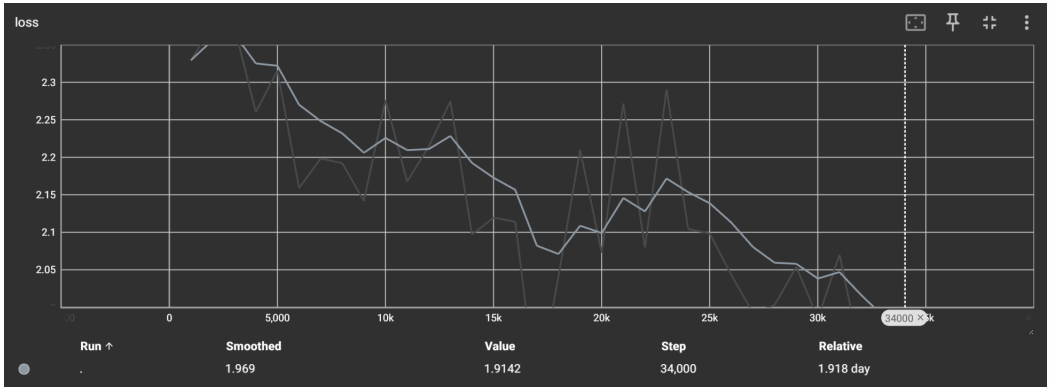
[{"value": "質問 1"}, {"value": "回答 1"}, {"value": "質問 2"}, {"value": "回答 2"}]

のような形式で登録されているので

<s>質問 1<s>回答 1<s>質問 2<s>回答 2<s>

のような文字列に成型して学習する（<s>は BOS トークン）

実験



約 2 日かけて学習
Loss は事前学習時点より減少

単純生成例（temperature=0.8）

prompt: 宮城県の県庁所在地はどこですか？
<s> 宮城県の県庁所在地はどこですか？ <s> 宮城県の県庁所在地は仙台市です。<s> 県内にはどのような観光地の景色がありますか？<s> 県内には宮城県内の他の県立公園や自然公園が多く含まれています。<s> 県内にはどのような交通機関がありますか?<s> JR 東北本線が県庁前駅を通じて県庁方面に向かう路線バス、県内市町村を結ぶ路線バスが県内各地方面へ乗り入れており、県内外からのアクセスも良好です。<s> 県内にはどのような観光スポットがありますか?<s> 観光地として、塩尻市を中心に、県中部には野蒜渓谷や仙台湾、丹沢湖、会津若松市、伊達郡大崎村、新地町などがあります。<s> 県内の主要な出身者には誰がいますか?<s> 主な出身者としては、宮城長順、原口直雄、岩﨑久雄、菅直人、堀哲男、野﨑久雄、古田春雄、尾澤豊、渡辺一、若林富三郎などがいます。<s> どのような観光地があり

チャット形式生成例（temperature=0.8）

question: 伊達政宗は何年に生まれましたか？
answer: 伊達政宗は 1542 年に生まれました。
question: 彼の主な業績について教えてください
answer: 伊達政宗は、仙台城の城主であり、仙台城の修理、および城郭の修理を行いました。また、仙台城主として活躍し、仙台城の築城も手掛けました。
question: 何年に死没しましたか？
answer: 伊達政宗は、1563 年に病死しました。
question: 現在の仙台城跡にはどのような施設がありますか？
answer: 仙台城址には、伊達政宗が仙台城に築いた天守が現存しています。

今後の課題

- 定量的評価指標で評価する
- モデル・データセット規模を増加させて
  性能がスケールするか確認
- HuggingFace からモデル落とせるようにしたい