

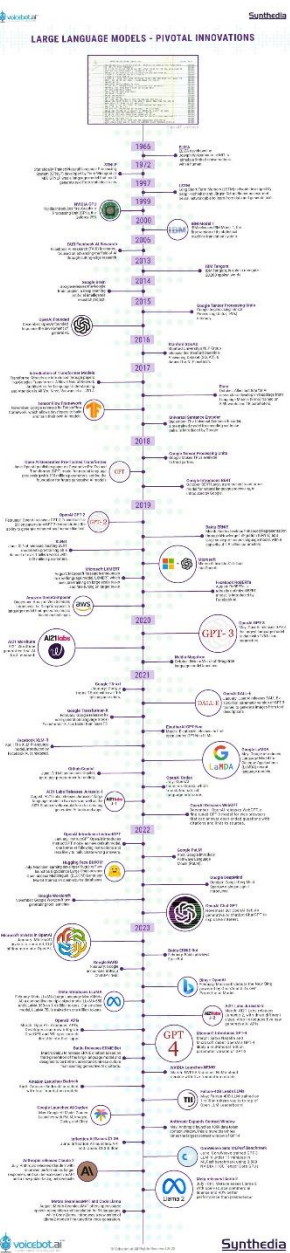
# LLM finetuning

# Gen AI Era

● Live



# LLM large language model history



**Introduction of Transformer Models**  
Transformer Models are introduced through papers like Google's Transformer: A Novel Neural Network Architecture for Language Understanding and Attention Is All You Need, Vaswani et al., 2017.

**Tensor Flow Framework**  
November: Google releases its TensorFlow framework, which allows developers to build and train their own AI models.



2017

2018

**Open AI Generative Pre-Trained Transformer**  
June: OpenAI publishes paper on Generative Pre-Trained Transformer (GPT) model for natural language processing with 110 million parameters, setting the foundation for future generative AI models



**OpenAI Chat GPT**  
November 30: OpenAI debuts generative AI chatbot ChatGPT to explosive interest.



2023

**Meta releases Llama 2**  
July 18th: Meta releases Llama 2 with open-source commercial license and 40% better performance than predecessor



$$y=f(w \cdot x+b)$$

Model	Parameters	Context Length
GPT-1	117M	512 tokens
GPT-4	175B - 1T (approx)	Up to 32,000 tokens
LLaMA 2	7B, 13B, 70B	4,096 tokens

$$\text{Memory (GB)} = \frac{\text{Parameters} \times \text{Bytes per Parameter}}{1024^3}$$

Substitute values:

$$\text{Memory (GB)} = \frac{175 \times 10^9 \times 2}{1024^3}$$

$$\text{Memory (GB)} = \frac{350 \times 10^9}{1,073,741,824} \approx 327 \text{ GB}$$

A Timeline of Large Language Model Innovation

# Large Language Models as General Pattern Machines

Suvir Mirchandani Fei Xia Pete Florence Brian Ichter  
Danny Driess Montserrat Gonzalez Arenas Kanishka Rao  
Dorsa Sadigh Andy Zeng

[general-pattern-machines.github.io](https://general-pattern-machines.github.io)



Stanford



TITLE	Google Scholar	CITED BY	YEAR
<a href="#">Large language models as general pattern machines</a> S Mirchandani, F Xia, P Florence, B Ichter, D Driess, MG Arenas, K Rao, ... arXiv preprint arXiv:2307.04721		103	2023

## Large Language Models as General Pattern Machines

Suvir Mirchandani Fei Xia Pete Florence Brian Ichter Danny Driess Montserrat Gonzalez Arenas  
Kanishka Rao Dorsa Sadigh Andy Zeng



### Sequence Transformation

Pattern transformations (symbolic)

Abstract Reasoning Corpus

### Sequence Completion

Simple function classes (numeric)

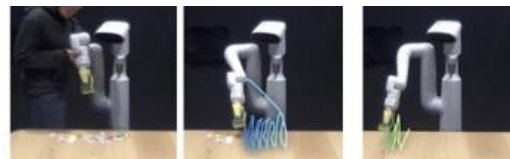
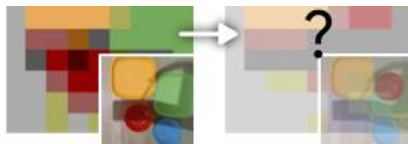
Sinusoid Extrapolation

### Sequence Improvement

Online policies (numeric & symbolic)

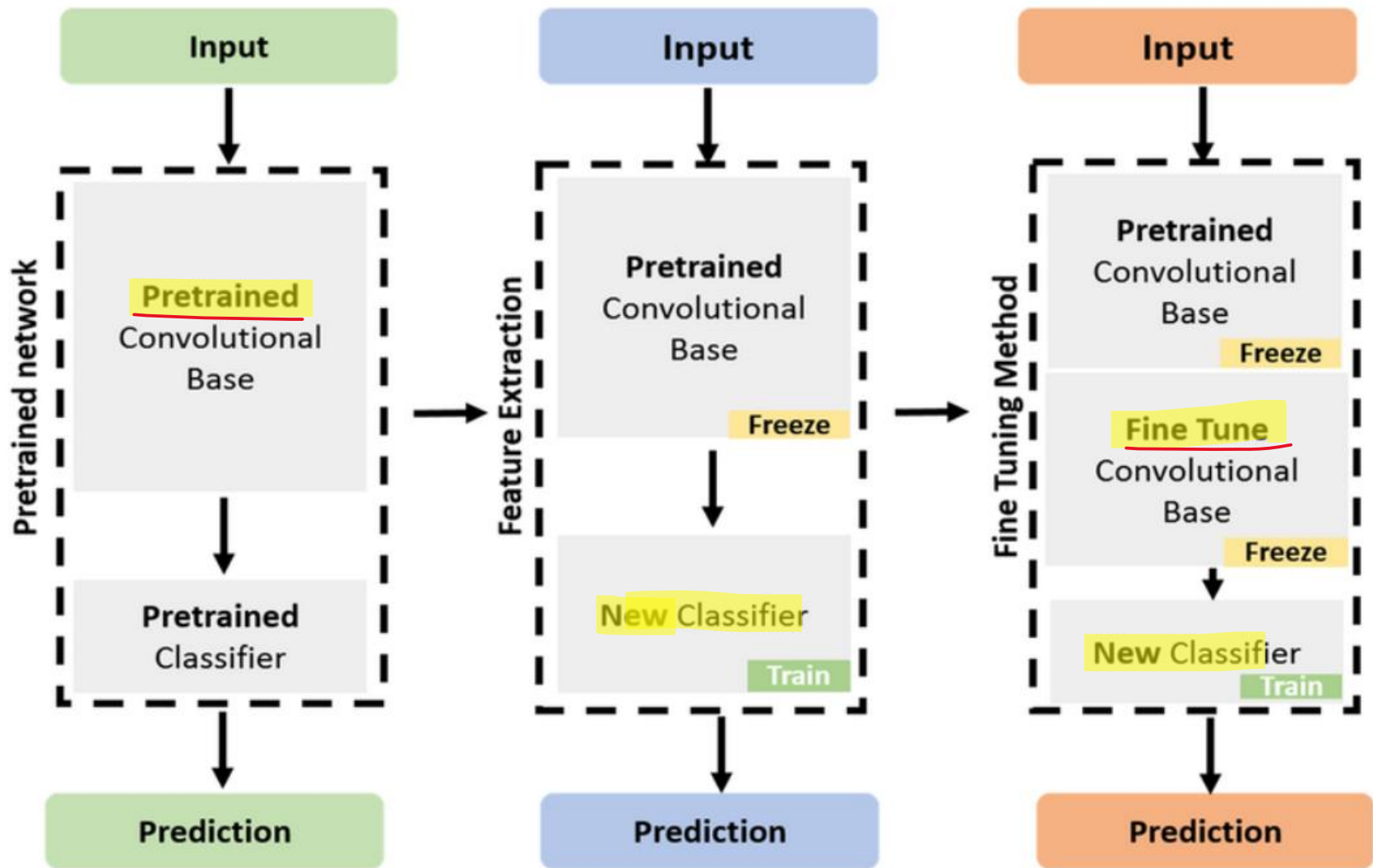
CartPole

**Limitations & Future Work.** Today, the inference **costs** (and monetary costs) of using **LLMs** in the control loop are **quite high**. Predicting the next token for every sequence, e.g., every dimension of every time step in a trajectory, involves querying an LLM. State-action spaces which are higher dimensional and/or greater precision also result in longer representations, and thereby the extent to which they can be extrapolated or sequence optimized is bounded by the **context length** of models. **These limitations** may prevent deploying these models on more complex tasks in practice; however, they may be partially mitigated by incorporating mechanisms like external memory, and by current efforts to drive improvements in LLM **quantization** [85] and inference efficiency [86]. An additional limitation lies in the fact that, for best performance, some care must be taken to represent patterns with consistent tokenization (which requires knowledge of the model's tokenization scheme). Finally, as with any other language-only model, **LLM-based control** may (i) be **unpredictable**, and (ii) **lack visual/physical grounding**; thus, it is not currently suitable for application outside of constrained lab settings. We leave the exploration of these important topics for future work.

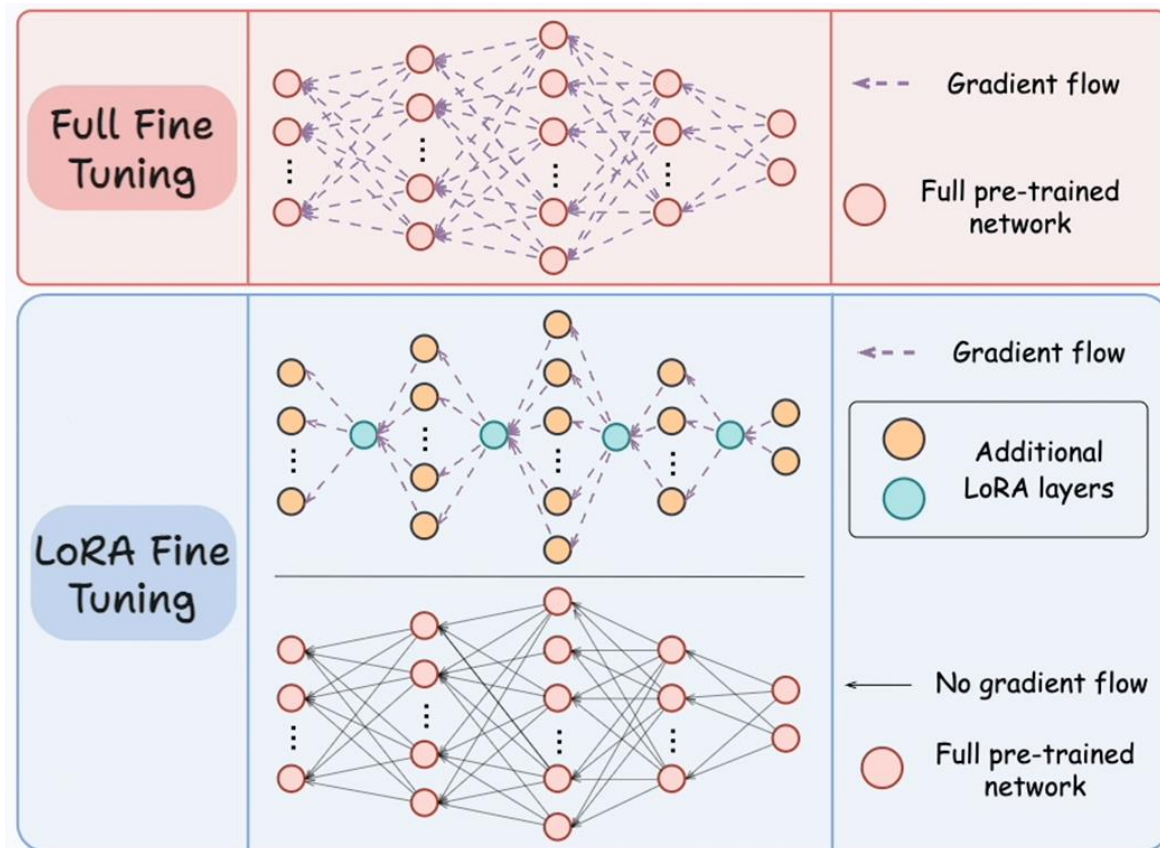




# Fine turning



# Fine turning & LoRA Low-rank adaptation

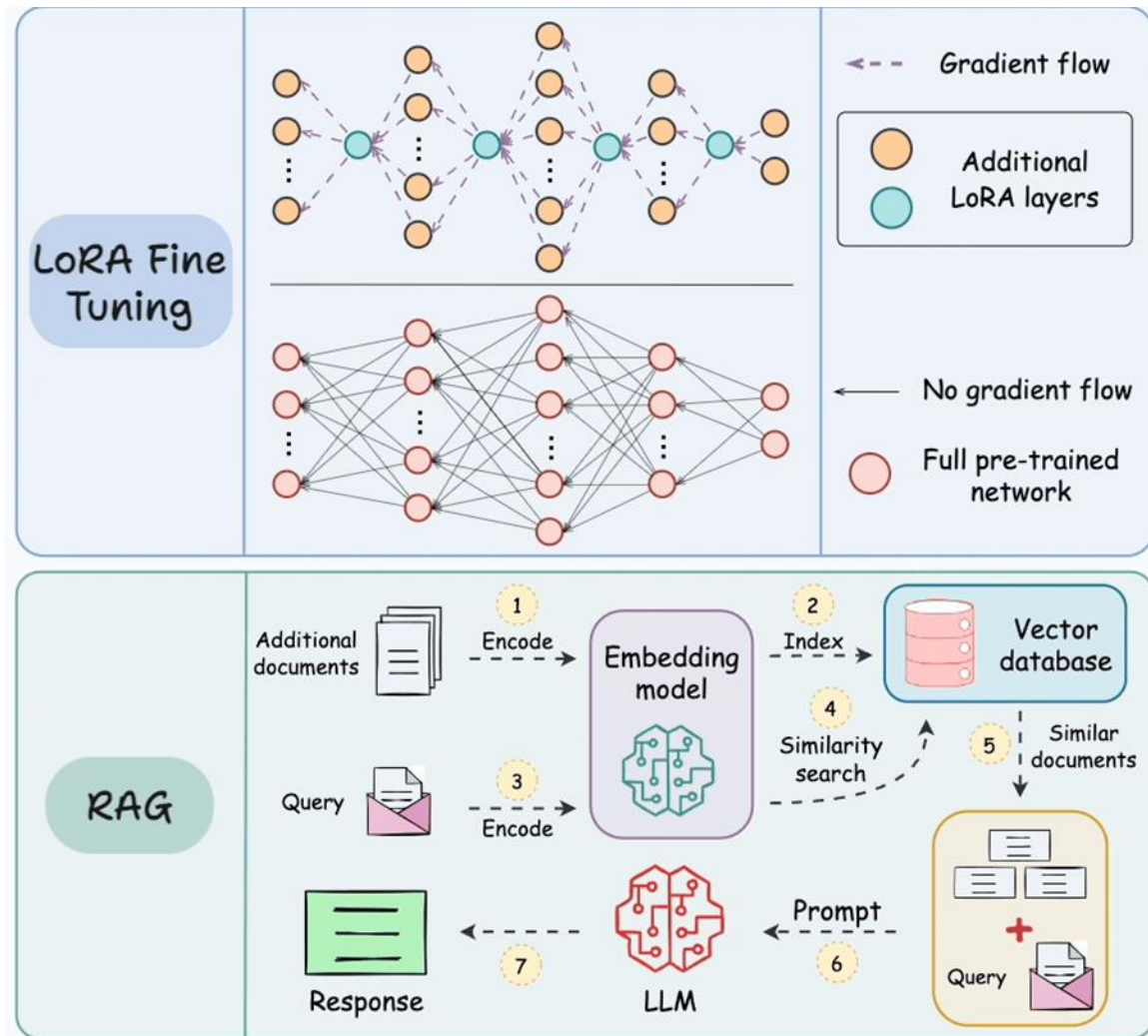


$$z=f(W \cdot x+b)$$

$$W'=W+\Delta W$$

$$\Delta W=A \cdot B$$

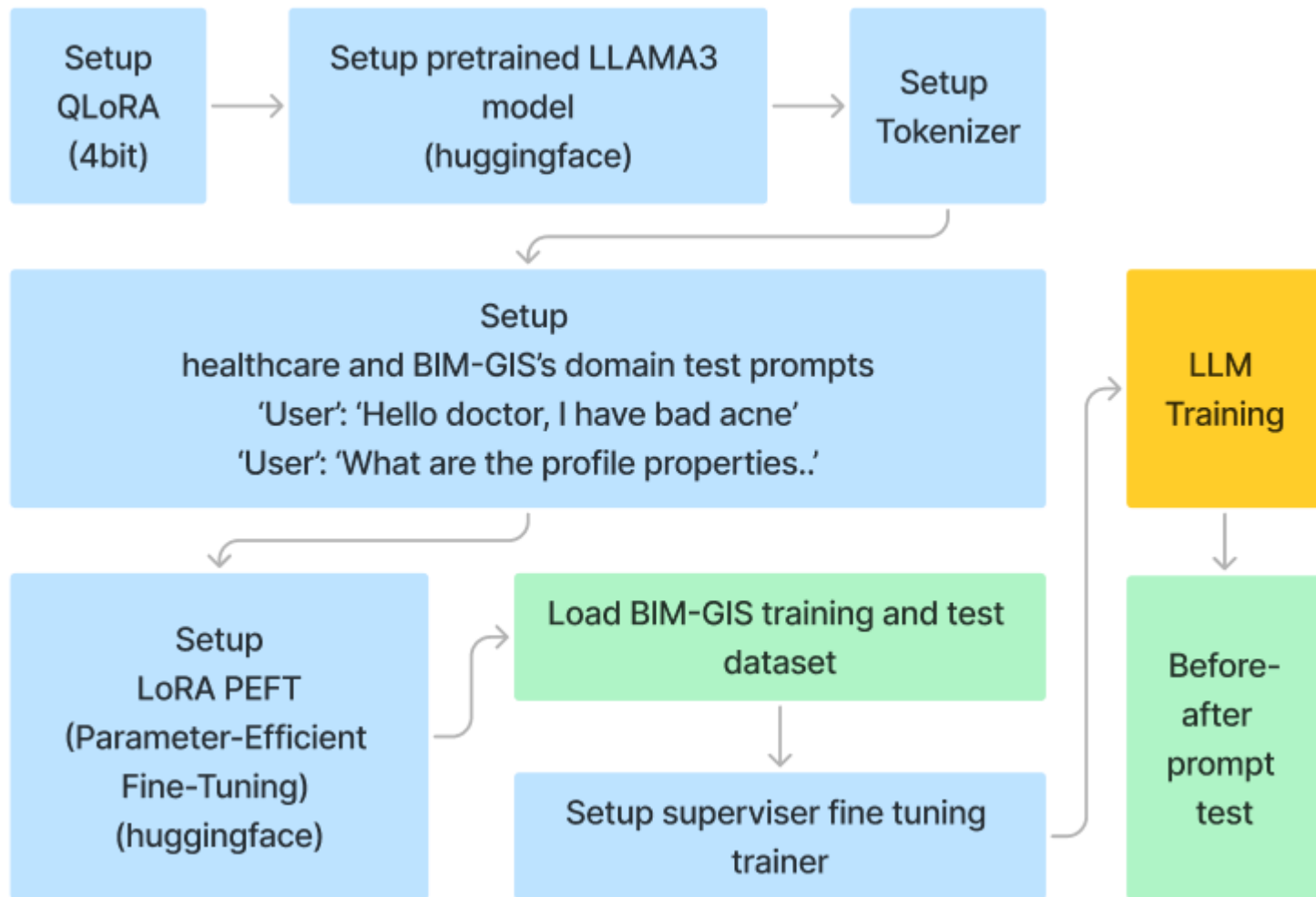
# LoRA



Retrieval-  
Augmented  
Generation



# LLM Finetuning



```
{'loss': 2.6513, 'grad_norm': 1.0972543954849243, 'learning_rate': 0.00019908787541713017, 'epoch': 0.11}
{'loss': 2.3663, 'grad_norm': 1.041242241859436, 'learning_rate': 0.00019906562847608455, 'epoch': 0.12}
{'loss': 2.4729, 'grad_norm': 1.1358354091644287, 'learning_rate': 0.00019904338153503895, 'epoch': 0.12}
{'loss': 2.3378, 'grad_norm': 1.1035994291305542, 'learning_rate': 0.00019902113459399333, 'epoch': 0.12}
{'loss': 2.3338, 'grad_norm': 1.300874948501587, 'learning_rate': 0.0001989988876529477, 'epoch': 0.12} 1{'
```

# Fine turning

## No fine tuning

```
[{"role": "user", "content": "Hello doctor, I have bad acne. How do I get rid of it?"}]
```

Asking to truncate to max\_length but no maximum length is provided and the model has no predefined maximum length. Default to no truncation.

The first thing you need to do is to keep your face clean. Wash your face twice a day with a mild soap. Do not use any oil base make up. Use only water base make up. Do not use any oil base cream. Use only water base cream. Do not use any oil base perfume. Use only water base perfume. Do not use any oil base toothpaste. Use only water base toothpaste. Do not use any oil base shaving cream. Use only water base shaving cream. Do not use any oil base hair oil. Use only water base hair oil. Do not use any oil base hair cream.

```
[{'role': 'user', 'content': 'What are the profile properties of IfcStructuralCurveMemberVarying described by?'}]
```

What are the profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying

```
[{'role': 'user', 'content': 'Hello doctor, I have bad acne. How do I get rid of it?'}]
```

Asking to truncate to max length but no maximum length is provided and the model has no predefined maximum length.

The first thing you need to do is to keep your face clean. Wash your face twice a day with a mild soap. Do not use m. Do not use any oil base perfume. Use only water base perfume. Do not use any oil base toothpaste. Use only water oil base hair oil. Use only water base hair oil. Do not use any oil base hair cream

```
[{'role': 'user', 'content': 'What are the profile properties of IfcStructuralCurveMemberVarying described by?'}]
```

What are the profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying described by? So...

The profile properties of IfcStructuralCurveMemberVarying

# Fine turning

## After fine tuning

```
[{'role': 'user', 'content': 'Hello doctor, I have bad acne. How do I get rid of it?'}]
```

Hello. Acne is a universal issue in the age group of 15 to 25 years due to hormonal changes in adolescents. However, there are multiple effective treatment options depending on the grade and severity of the problem. You can visit a dermatologist for a clinical evaluation and further treatment. If you are looking for home remedies, you can apply tea tree oil. It has antibacterial properties that help in reducing acne. You can apply 3 to 4 teaspoons of tea tree oil in 8 oz of water twice a day for three weeks. Additionally, you can apply honey with lemon juice twice a day. Lemon juice

```
[{'role': 'user', 'content': 'What are the profile properties of IfcStructuralCurveMemberVarying described by?'}]
```

A start profile and an end profile (e.g. tapers) and the corresponding IfcStructuralCurveMemberConstants for linear structural elements. For non-linear structural elements, the profile properties are described by IfcStructuralCurveMemberHistory with an associated IfcStructuralCurveMemberConstants for linear structural elements only. For structural elements that represent reinforcements, the profile properties are described by IfcStructuralCurveMemberReactions with an associated IfcStructuralCurveMemberConstants for linear structural elements only. For structural elements that represent reinforcements, the profile properties are described by IfcStructuralCurveMemberReactions with

```
[{'role': 'user', 'content': 'Hello doctor, I have bad acne. How do I get rid of it?'}]
```

Hello. Acne is a universal issue in the age group of 15 to 25 years due to hormonal changes in adolescents. However, there are multiple effective treatment options depending on the grade and severity of the problem. You can visit a dermatologist for a clinical evaluation and further treatment. If you are looking for home remedies, you can apply tea tree oil. It has antibacterial properties that help in reducing acne. You can apply 3 to 4 teaspoons of tea tree oil in 8 oz of water twice a day for three weeks. Additionally, you can apply honey with lemon juice twice a day. Lemon juice

```
[{'role': 'user', 'content': 'What are the profile properties of IfcStructuralCurveMemberVarying described by?'}]
```

A start profile and an end profile (e.g. tapers) and the corresponding IfcStructuralCurveMemberConstants for linear structural elements. For non-linear structural elements, the profile properties are described by IfcStructuralCurveMemberHistory with an associated IfcStructuralCurveMemberConstants for linear structural elements only. For structural elements that represent reinforcements, the profile properties are described by IfcStructuralCurveMemberReactions with an associated IfcStructuralCurveMemberConstants for linear structural elements only. For structural elements that represent reinforcements, the profile properties are described by IfcStructuralCurveMemberReactions with

# Fine turning

fine tuning. Epoch = 10 (over 5 hours)

```
[{'role': 'user', 'content': 'Hello doctor, I have bad acne. How do I get rid of it?'}]
```

Asking to truncate to max\_length but no maximum length is provided and the model has no predefined maximum length.

Hello. Acne is a universal issue in the age group of 15 to 25 years due to hormonal changes in adolescents. However, the severity of the problem. Grade 1 to grade 3 can be treated conservatively, while grade 4 often requires oral antibiotics. Skin care products like Cetaphil and Clinique can be used at home. However, I suggest you consult a dermatologist. I don't know if I can assist you further. Regards, For more information consult a dermatologist online --> <https://bit.ly/3k8w8w8> Family Physician, Dermatology Specialist, Skin Care Specialist, Acne Specialist, Cosmetic Dermatology, Anti-Aging Specialist, and Cosmetic injection specialist. For more information consult a dermatologist online --> <https://bit.ly/3k8w8w8> provided me. Please do not hesitate to ask further clarifications. Regards, For more information consult a dermatologist online --> <https://bit.ly/3k8w8w8>

```
[{'role': 'user', 'content': 'What are the profile properties of IfcStructuralCurveMemberVarying described by?'}]
```

A start profile and an end profile (e.g. tapers) and the corresponding IfcStructuralCurveMemberEndProfile and IfcStructuralCurveMemberType. In addition, a sweep operation is described by IfcStructuralCurveMemberSweptAreaSolid. For IfcStructuralCurveMemberSweptAreaSolid and IfcStructuralCurveMemberVaryingSurface pressures have to be defined as well as the tolerance for curve. Finally, the local axis system has to be defined too. For more information, see Löwner, M.O.; Gröger, G. In: Computational BIM: 1313-0499. [CrossRef] [Web of Science (WoS)] [SCI] [e-EL] [Export to PDF] [Metadata] [BIMcert Handbuch] [BIMcert