

DeepSeek-OCR

Heng-Shiou Sheu
2025 Oct.

Why DeepSeek-OCR Went Viral

Frequently discussed on social media for its concept of Optical Compression.



DeepSeek-OCR: Contexts Optical Compression

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

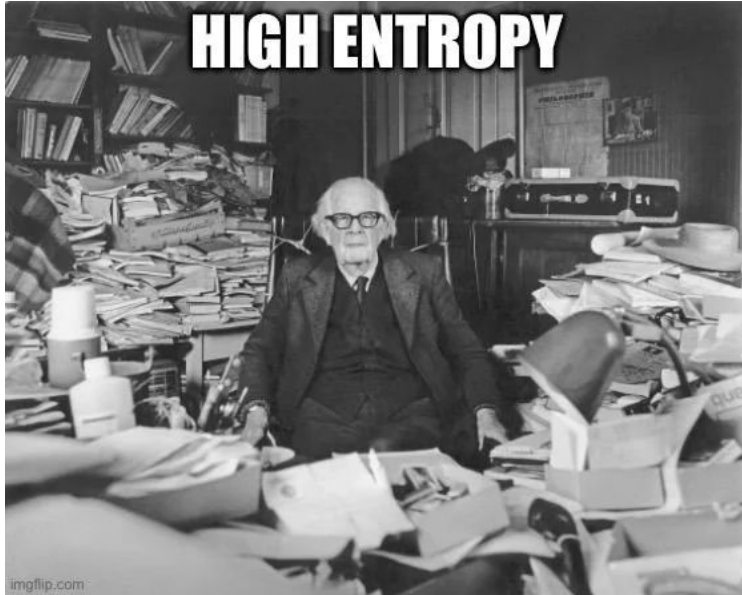


Andrej Karpathy  @karpathy · Oct 21

I quite like the new [DeepSeek-OCR paper](#). It's a good OCR model (maybe a bit worse than dots), and yes data collection etc., but anyway it doesn't matter.

Information Entropy

How humans compress infinite meanings into finite symbols



船

boat

舟 八 口

vessel eight people

The Discovery of Genesis, C.H. Kong and Ethel Nelson, p. 55

Language is compression.

Tokenizer

It's how symbols begin to imitate meaning

Table 1: Tokenizer comparisons between original LLaMA and Chinese LLaMA.

	Length	Content
Original Sentence	28	人工智能是计算机科学、心理学、哲学等学科融合的交叉学科。
Original Tokenizer	35	‘人’，‘工’，‘智’，‘能’，‘是’，‘计’，‘算’，‘机’，‘科’，‘学’，‘、’，‘心’， ‘理’，‘学’，‘、’，‘0xE5’，‘0x93’，‘0xB2’，‘学’，‘等’，‘学’，‘科’，‘0xE8’， ‘0x9E’，‘0x8D’，‘合’，‘的’，‘交’，‘0xE5’，‘0x8F’，‘0x89’，‘学’，‘科’，‘。’
Chinese Tokenizer	16	‘人’，‘工’，‘智’，‘能’，‘是’，‘计’，‘算’，‘机’，‘科’，‘学’，‘、’，‘心’，‘理’，‘学’，‘、’，‘哲’，‘学’， ‘等’，‘学’，‘科’，‘融’，‘合’，‘的’，‘交’，‘叉’，‘学’，‘科’，‘。’

A balance between structure and understanding.

Compression

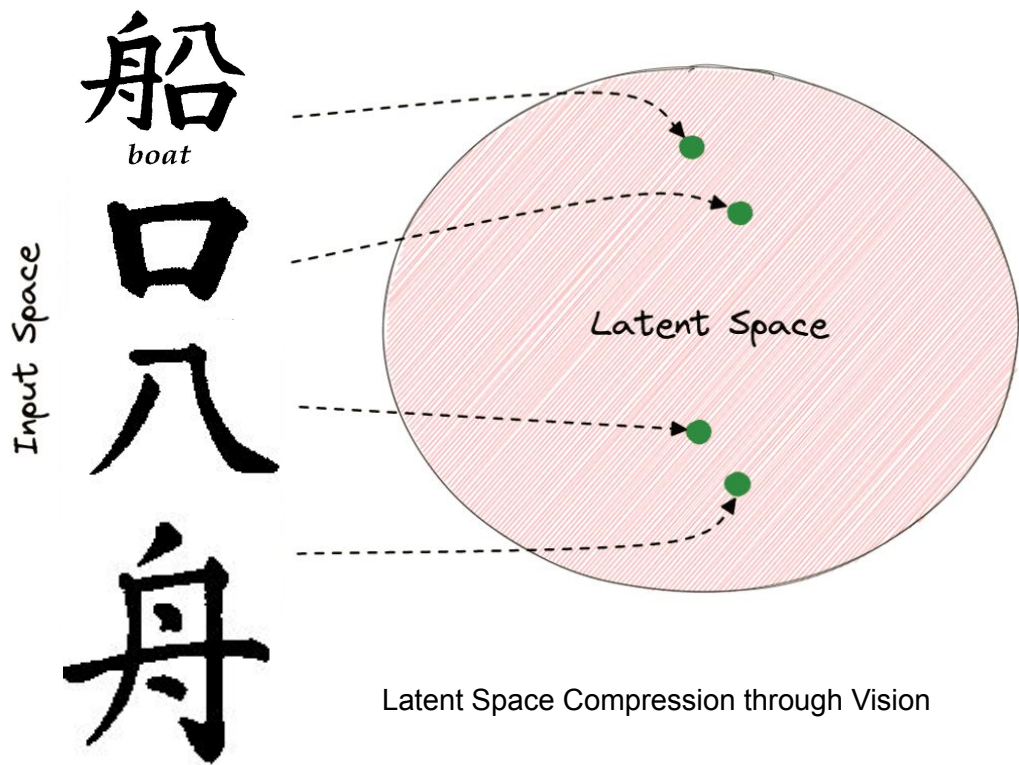
What happens when we move beyond text?

Text Tokens	Vision Tokens =64		Vision Tokens=100		Pages
	Precision	Compression	Precision	Compression	
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

Exploring the mapping between visual tokens and text tokens.

Beyond the Limits of Language Compression

What if we could encode meaning—not in words—but in the latent space of vision?



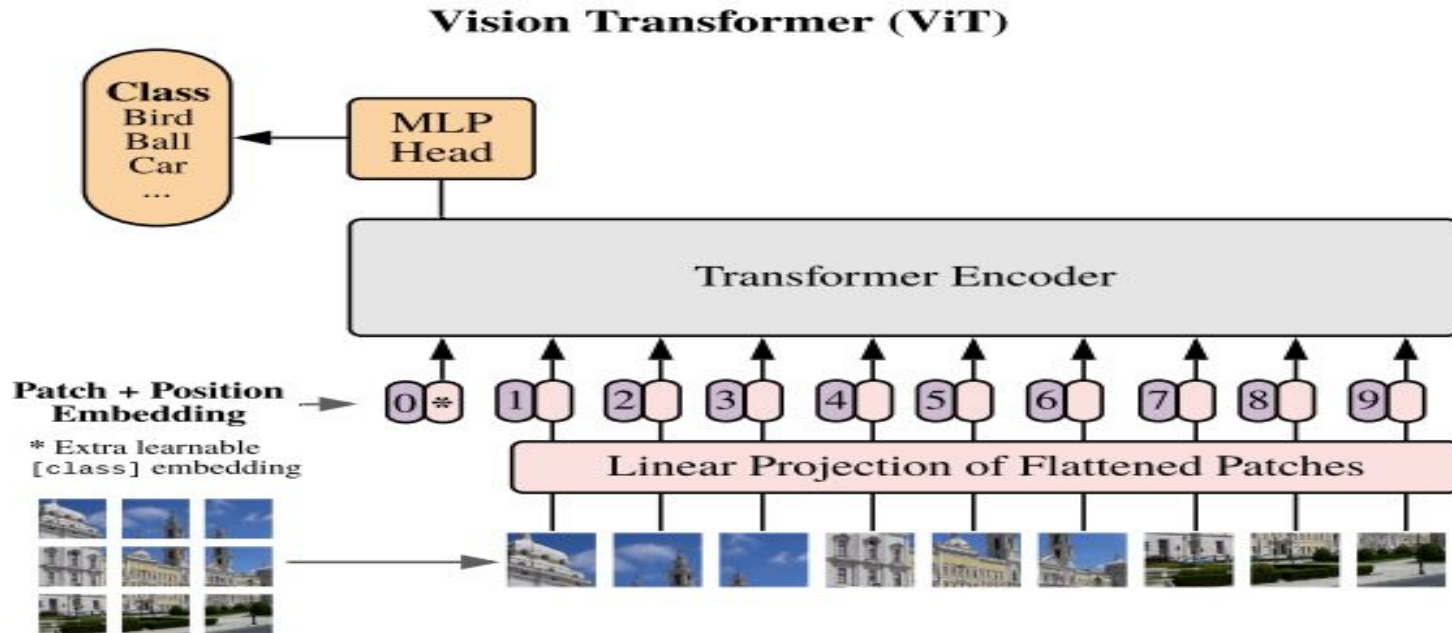
How can a model “see” an image?

Vision Tokens — the bridge connecting visual input and language understanding



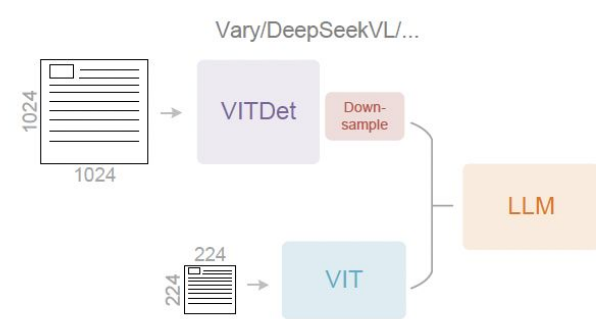
How can a model “see” an image?

Vision Tokens — the bridge connecting visual input and language understanding

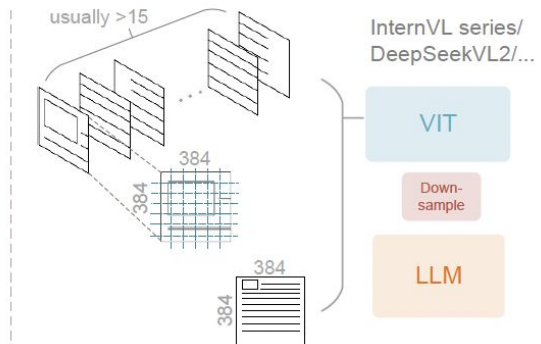


The Trade-off of Vision Tokens

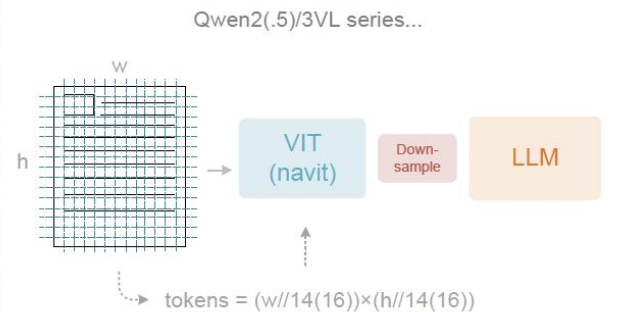
Sequence Explosion and the Compression Bottleneck



- [x] unsupported pipeline parallel
- [x] unsupported extreme resolution
- [x] two pre-processes
- [x] hard to deployment



- [x] low native resolution
- [x] too many vision tokens
- [x] overly small patches
- [x] small global view



- [x] too many vision tokens
- [x] large activations
- [x] need long sequence length
- [x] slow inference speed

Fewer tokens. Same meaning.

A two-stage compression architecture

A two-stage compression framework that enables visual information to retain its full semantic meaning with the fewest possible tokens.

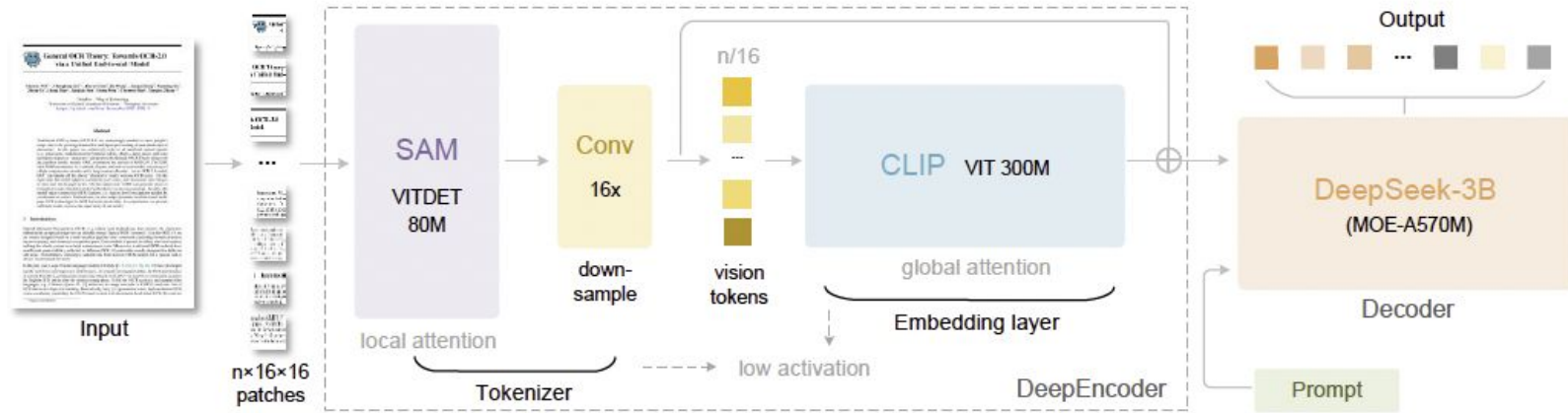
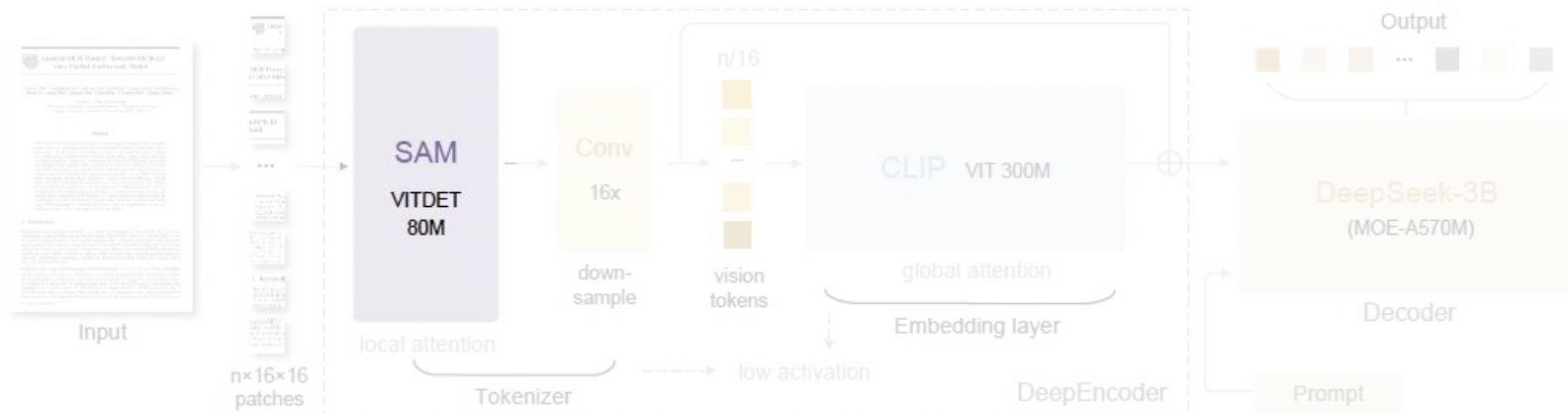


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three keeping meaning intact, even with minimal visual tokens.

DeepEncoder

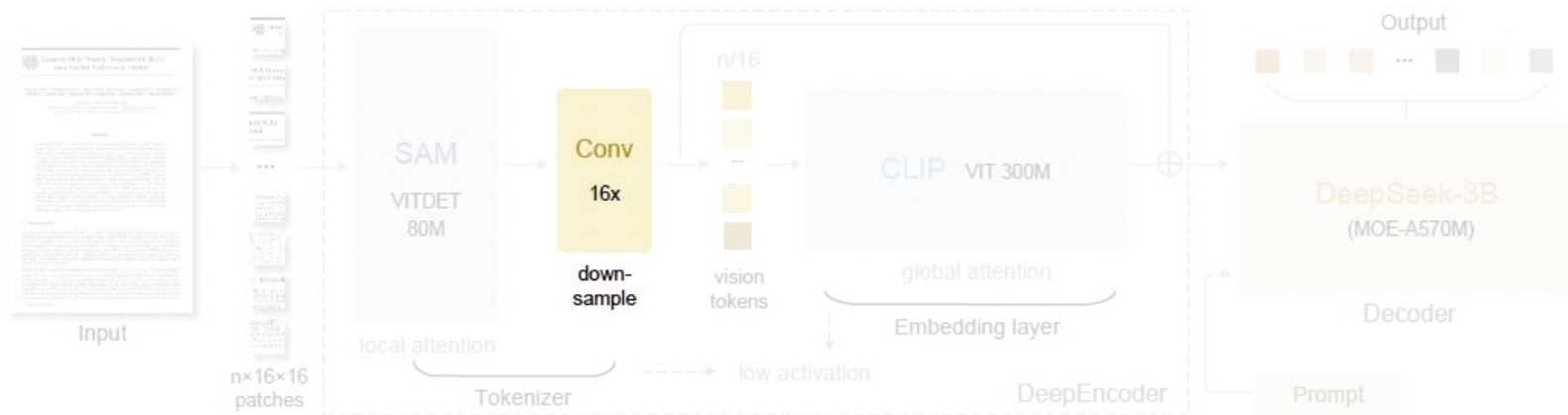
Local Information: Detail Extraction with SAM-base



- Windowed Attention → Focus on patch-level details
- 1024×1024 image → 16×16 patches
- 4096 Vision Tokens generated

DeepEncoder

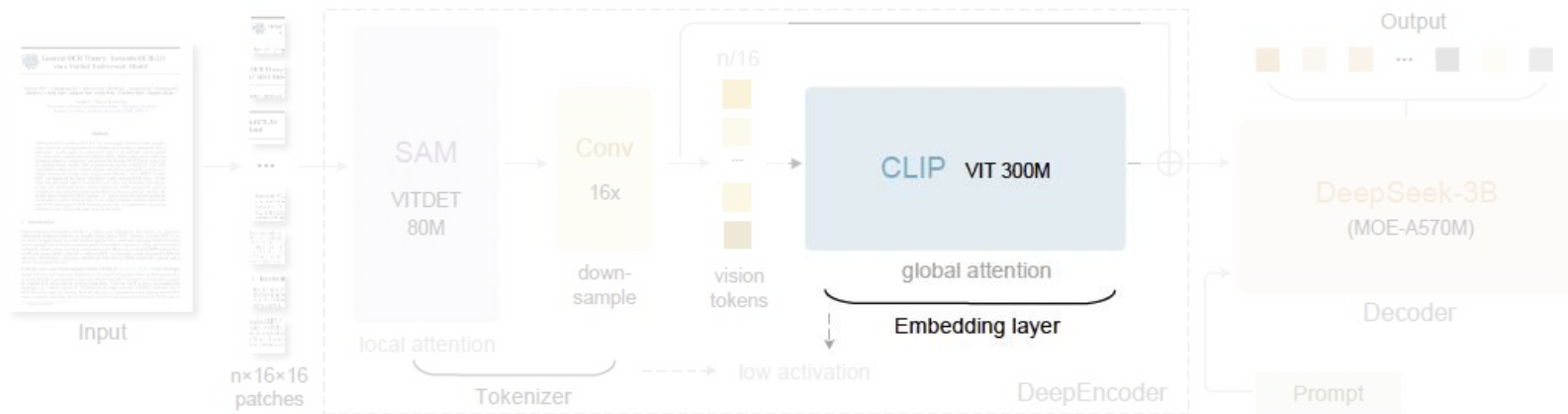
16× Convolutional Compressor: Efficient Dimensionality Reduction



- Using 3×3 kernels (stride 2, padding 1),
- it expands the feature channels from 256 to 1024,
- compressing 4096 patch tokens into just 256.

DeepEncoder

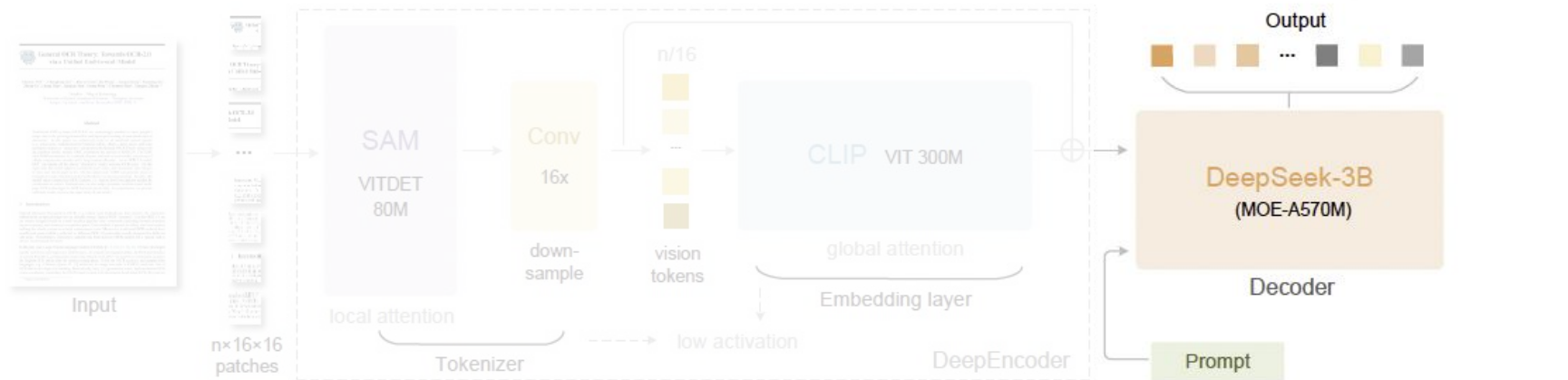
Global Semantic Integration (CLIP-Large)



- Extracts global semantics & context
- Removes patch embedding layer
- Dense global attention for holistic meaning

DeepDecoder

Achieving large-model expressiveness with small-model effic

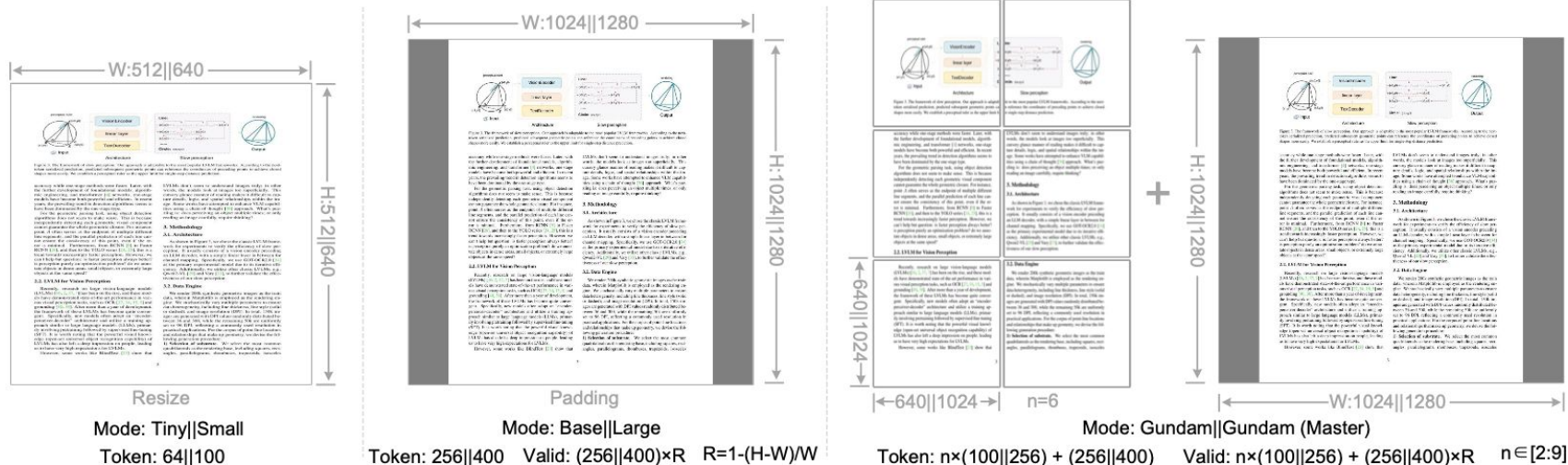


- Built on the **DeepSeekMoE-3B-MoE** architecture
- Activates **6 routing experts** and **2 shared experts**, for a total of approximately **570 million active parameters**

$$f_{\text{dec}} : \mathbb{R}^{n \times d_{\text{latent}}} \rightarrow \mathbb{R}^{N \times d_{\text{text}}}; \quad \hat{\mathbf{X}} = f_{\text{dec}}(\mathbf{Z}) \quad \text{where } n \leq N$$

Multi-Resolution Capability

The DeepSeek-OCR model supports multiple input resolutions.



- Tiny(512×512, 64 token),
- Small(640×640, 100 token),
- Base(1024×1024, 256 token),
- Large(1280×1280, 400 token)

- Dynamic Modes: *Gundam / Gundam-Master*
- Ultra-high-resolution inputs
- Tiling strategy
- Local details + Global context

Real-World Application Scenarios

Traditional OCR Use Case: Based on the OCR1 dataset

<image>\nFree OCR.

نور وكالات الدعم في إنشاء المؤسسات الصغيرة والمتوسطة لتوفير مناصب الشغل

- [illegible]

II - الطريقة والأدوات :

(1) مجتمع وعينة الدراسة

استهدفت الدراسة عين عتوائية قدرت بـ 83 فرد. حيث وزعت عليهم استمارات الاستبانة ولم اسراجها كلها (83 استبانة) أي ما يعادل 100% كذلك تم تكن هناك استبانات مغلقة وهذا أدى على حداثة التحيوت في الإجابة عليها واستكمالها لشروط مكنها. وبالتالي يمكن الاستمارات للوزعة كانت صالحة وقابلة للتحليل الإحصائي، ويمكن تلخيص ما سبق في الجدول التالي:

الجدول رقم (01): عدد الإحصائيات الموزعة والمستلمة		
الاستمارات	العدد	النسبة (%)
الموزعة	83	100
غير مسترجعة	00	00
إضافة للفحليل	83	100

المصدر: من إعداد الباحث

(2) معيّنات الدراسة

من خلال إشكالية البحث ووفق الدراسة التطبيقية لتحديد لنا متغيرات الدراسة في متغيرين أحدهما مستقل والآخر تابع.

فمتغير الدراسة المستقل يتحلى بـ: وكالات الدعم.

أما متغير الدراسة التابع فيتمثل في: إنشاء مشاريع صغيرة ومتوسطة لتطوير مناصب الشغل.

(3) بناء الاستبانة: لقد تم إعداد الاستبانة بشكل يلد أهداف الدراسة وفق الفرضيات المقترحة حيث تضمن أولاً المعلومات الشخصية للعبة وذلك التعرف على خصائصها، ثم تطرقوا إلى أسئلة حول موضوع البحث ومتغيرات الإشكالية، وقد تم تبين الشكل المطلق في تصميم

زور وكالات الدعم في إنشاء المؤسسات الصغيرة والمتوسطة لتوفير مناصب الشغل

4. الكلفة الوطنية لتسيير الحضرية JANGEM التابعة بموجب المرسوم الصادر رقم 14/04 المؤرخ 14/04/2012 على 22 يناير 2012، تغطي تدابير علاج عاجل لتسليم المستشفيات لمرضى الأمراض المعدية في ظل تفشي فيروس نقص المناعة البشرية، وتضمنت الإجراءات الوقائية الممنوعة في الكشافة هذه؛
تسهيل المبادئ من أجل الحماية والحفاظ على الصحة، وتنفيذ المبادرات الممنوعة في الكشافة هذه؛
الجهاز في الصندوق الوطني لضمان العمل في الظلابة أو تحديد النشاط الاجتماعي التابع لإدارة العمل والصحة المهنية، والصحة المهنية.
 5. مهام الكلية: تطبيق سياسة الدولة في مجال حماية البيئة والمخاطر عن طريق تدعيم المبادرات الفردية في كل من مساهمته على نطاقات لصياغة القرار وتنفيذ دور الكلية بغير التمييز بين الاستشارة والمراقبة والمراقبة والتقييم وضمان الصيانة لإنتاج المشاريع (المشاريع 2004).
 6. الضرر والصحة عن غيرة عن طريق فصل الـ 500000 حوزة موجهة لفتح الطالين والمنازل التي تلغوا من 18 سنة فما فوق ومكتسبين أيضا وأحرف في مناطق ضيقة.
 7. الصندوق الوطني للتأمين على البطالة (CNAO) استحدثت هذه الهيئة منذ 2004. ويهدف هذا المرسوم (1994) على الدورات إعادة الانخراط في الحياة المهنية والمساهمة في تمويل أساسيين (2004).
- تتم التأسيس على البطالة
- جهاز دعم استحداث النشاطات من طرف البطالين عن المشاريع والتعاون المراهق من سنة 2003 في
- صندوق ضمان البطالة
- يضمن استمرار بطالة البطالين والمخاطر عن طريق تحميل النشاطات المخصصة لتحويل الـ التأمين عن البطالة ورعاية الكه والضعاف
- يساعد و يدعم بالتواصل مع المصالح العمومية لتحويل الـ التأمين عن البطالة
- يرأس ويخطط صندوق الحفاظ على ممتلكاته من مواجهة التزاحم تجاه المستفيدين في جميع الظروف
- الطريقه بالولايات II -

1. مجتمع وعينة الدراسة استهدفت الدراسة عينة عشوائية قدرت بـ 83 فرد. حيث وزعت عليهم استمارات الاستبانة وتم استرجاعها كلها (83 استمارة) أي ما يعادل 100% ككل لم تكن هناك استبانات ملغاة وهذا دليل على جدية المبحوث في الإجابة واستكملها لنصير الموضوع وبالتالي فكل الاستمارات الموزعة كانت صالحة مقبلة للتحليل. تلخيصاً لما سبق في الجدول التالي:

الجدول رقم (01): عدد الاستبيانات الموزعة والمسئمة

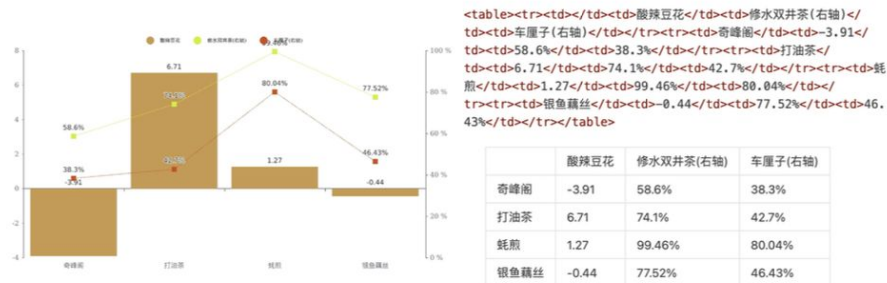
النسبة %	العدد	الاستمارات
100	83	الموزعة
0	0	الغير مسترجعة
100	83	الصالحة للتحليل

لمصدر: من إعداد الباحث

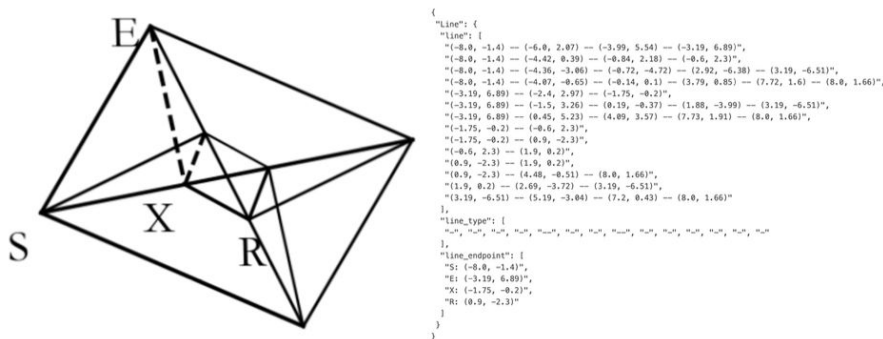
2. معقورات الدراسة من خلال إشكالية البحث ووفق الدراسة التطبيقية تتحدد لنا معقورات الدراسة في معقورين أساسيين هما: معقور التالى والمقور السابق، فمعقور الدراسة السابق يتجلى في: وكالات الدعم، أمّا معقور الدراسة التالى فيتمثل في: إنشاء مشاريع صغيرة وموسطة لتوفير مناصب الشغل.
3. بناء الاستمالية: لقد تم إعداد الاستبيان بشكل فردي وفق الفرضيات المقترحة حيث تضمن أول المعلومات الشخصية للعبنة وذلك لتعرف على خصائصها، ثم تطرقنا إلى استئدة كل موضوع (المقور السابق، ومعقورات الاستمالية، وقد تم تدوين النتائج الطلقة في تصميم

Real-World Application Scenarios

Advanced OCR Scenarios: OCR2 Dataset



(a) Image-text ground truth of chart



(b) Image-text ground truth of geometry

- Chart analysis: Converts visual charts to structured HTML table format
- Chemical formula recognition: Transforms molecular structures to SMILES notation
- Geometric parsing: Processes mathematical figures into structured coordinate systems
- Beyond what traditional OCR can do.

Real-World Application Scenarios

DeepSeek-OCR retains VLM-level generalization using the OCR3 dataset.

<image>\nLocate <ref>11-2</ref> in the image.



<image>\nIdentify all objects in the image and output them in bounding boxes.



<image>\nDescribe this image in detail.



<image>\nLocate <ref>the teacher</ref> in the image.



<image>\n这是一张



照片, 展示了一辆红色的消防栓。消防栓上有一个笑脸, 显得非常友好和亲切。消防栓的顶部有一个黑色的盖子, 周围有一些金属铆钉。在消防栓的底部, 有一个粉红色的贴纸, 上面写着 "bowtie"。背景中可以看到一条街道, 街道上有几辆停放的汽车和一些树木。整体画面给人一种温馨和友好的感觉。

君不见，黄河之水天上来

，奔流到海不复回。君不见，高堂明镜悲白发，朝如青丝暮成雪。人生得意须尽欢，莫使金樽空对月。天生我材必有用，千金散尽还复来。烹羊宰牛且为乐，会须一饮三百杯。岑夫子，丹丘生，将进酒，杯莫停。与君歌一曲，请君为我倾耳听。钟鼓馔玉不足贵，但愿长醉不愿醒。古来圣贤皆寂寞，惟有饮者留其名。陈王昔时宴平乐，斗酒十千恣欢谑。主人何为言少钱，径须沽取对君酌。五花马，千金裘，呼儿将出换美酒，与尔同销万古愁。

Performance Evaluation

OCR Task Results: Achieving Efficiency Beyond SOTA

Model	Tokens	English					Chinese				
		overall	text	formula	table	order	overall	text	formula	table	order
Pipeline Models											
Dolphin [11]	-	0.356	0.352	0.465	0.258	0.35	0.44	0.44	0.604	0.367	0.351
Marker [1]	-	0.296	0.085	0.374	0.609	0.116	0.497	0.293	0.688	0.678	0.329
Mathpix [2]	-	0.191	0.105	0.306	0.243	0.108	0.364	0.381	0.454	0.32	0.30
MinerU-2.1.1 [34]	-	0.162	0.072	0.313	0.166	0.097	0.244	0.111	0.581	0.15	0.136
MonkeyOCR-1.2B [18]	-	0.154	0.062	0.295	0.164	0.094	0.263	0.179	0.464	0.168	0.243
PPstructure-v3 [9]	-	0.152	0.073	0.295	0.162	0.077	0.223	0.136	0.535	0.111	0.11
End-to-end Models											
Nougat [6]	2352	0.452	0.365	0.488	0.572	0.382	0.973	0.998	0.941	1.00	0.954
SmolDocling [25]	392	0.493	0.262	0.753	0.729	0.227	0.816	0.838	0.997	0.907	0.522
InternVL2-76B [8]	6790	0.44	0.353	0.543	0.547	0.317	0.443	0.29	0.701	0.555	0.228
Qwen2.5-VL-7B [5]	3949	0.316	0.151	0.376	0.598	0.138	0.399	0.243	0.5	0.627	0.226
OLMOCR [28]	3949	0.326	0.097	0.455	0.608	0.145	0.469	0.293	0.655	0.652	0.277
GOT-OCR2.0 [38]	256	0.287	0.189	0.360	0.459	0.141	0.411	0.315	0.528	0.52	0.28
OCRFlux-3B [3]	3949	0.238	0.112	0.447	0.269	0.126	0.349	0.256	0.716	0.162	0.263
GPT4o [26]	-	0.233	0.144	0.425	0.234	0.128	0.399	0.409	0.606	0.329	0.251
InternVL3-78B [42]	6790	0.218	0.117	0.38	0.279	0.095	0.296	0.21	0.533	0.282	0.161
Qwen2.5-VL-72B [5]	3949	0.214	0.092	0.315	0.341	0.106	0.261	0.18	0.434	0.262	0.168
dots.ocr [30]	3949	0.182	0.137	0.320	0.166	0.182	0.261	0.229	0.468	0.160	0.261
Gemini2.5-Pro [4]	-	0.148	0.055	0.356	0.13	0.049	0.212	0.168	0.439	0.119	0.121
MinerU2.0 [34]	6790	0.133	0.045	0.273	0.15	0.066	0.238	0.115	0.506	0.209	0.122
dots.ocr ^{†200dpi} [30]	5545	0.125	0.032	0.329	0.099	0.04	0.16	0.066	0.416	0.092	0.067
DeepSeek-OCR (end2end)											
Tiny	64	0.386	0.373	0.469	0.422	0.283	0.361	0.307	0.635	0.266	0.236
Small	100	0.221	0.142	0.373	0.242	0.125	0.284	0.24	0.53	0.159	0.205
Base	256(182)	0.137	0.054	0.267	0.163	0.064	0.24	0.205	0.474	0.1	0.181
Large	400(285)	0.138	0.054	0.277	0.152	0.067	0.208	0.143	0.461	0.104	0.123
Gundam	795	0.127	0.043	0.269	0.134	0.062	0.181	0.097	0.432	0.089	0.103
Gundam-M ^{†200dpi}	1853	0.123	0.049	0.242	0.147	0.056	0.157	0.087	0.377	0.08	0.085

- Small mode: Outperforms GOT-OCR 2.0 using only 100 vision tokens (vs. 256).
- Gundam mode: Outperforms MinerU 2.0 with under 800 tokens (vs. ~7,000).

Tutorial

Github Colab

4.1 Free OCR (推薦)

```
: prompt = "
```

```
The attention mask and the pad token id were not set. As a consequence,
e pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:None for open-end generation.
```

```
=====
BASE: torch.Size([1, 256, 1280])
```

```
PATCHES: torch.Size([6, 100, 1280])
=====
```

```
| Period Ending | Dec 31, 2008 | Dec 31, 2009 | Dec 31, 2010 |
|-----|-----|-----|-----|
| **Assets** | | | |
| **Current Assets** | | | |
| Cash And Cash Equivalents | 8,656,672 | 10,198,000 | 13,630,000 |
| Short Term Investments | 7,189,099 | 14,287,000 | 21,345,000 |
| Net Receivables | 2,928,297 | 3,845,000 | 5,261,000 |
| Inventory | - | - | - |
| Other Current Assets | 1,404,114 | 837,000 | 1,326,000 |
| **Total Current Assets** | 20,178,182 | 29,167,000 | 41,562,000 |
| **Long Term Investments** | 85,160 | 129,000 | 523,000 |
| Property Plant and Equipment | 5,233,843 | 4,845,000 | 7,759,000 |
| Goodwill | 4,839,854 | 4,903,000 | 6,256,000 |
| Intangible Assets | 996,690 | 775,000 | 1,044,000 |
| Accumulated Amortization | - | - | - |
| Other Assets | 433,846 | 415,000 | 442,000 |
| Deferred Long Term Asset Charges | - | 263,000 | 265,000 |
| **Total Assets** | 31,767,575 | 40,497,000 | 57,851,000 |
```

- <https://github.com/Heng-xiu/all-things-llm/blob/main/talks/DeepSeek%20OCR.ipynb>