Simple Explanation

This research paper talks about a big machine translation competition that happened in 2024. The competition is called **WMT (Workshop on Machine Translation)** — it's the world's biggest yearly event for testing translation systems.

Here's what the paper says, in simple words:

---

1. Many teams built translation systems

Participants were asked to build machine translation systems for **11 different language pairs** (Example: English→German, English→Chinese, etc.)

---

2. Each system was tested on many domains

They didn't test systems on just one type of text.

They used **3 to 5 domains**, meaning different types of writing:

- News
- Medical
- Social media
- E-commerce
- Government text
  etc.

This is important because good translation should work everywhere — not just in one area.

---

3. They ALSO collected translations from many large language models (LLMs)

They didn't test only the participants' systems.

They ALSO tested translations from:

- **8 Large Language Models (LLMs)** like GPT, Claude, Llama, etc.
- **4 online translation tools** like Google Translate, DeepL, etc.

So they compared:

✔ Human-built systems
with
✔ Big AI models
with
✔ Online translators

---

4. They used a NEW way of evaluating translations

It's called **Error Span Annotation (ESA)**.

Let's explain this simply.

What is ESA? (SUPER SIMPLE)

Normally, translation quality is judged by:

- giving a score
- comparing with reference translations
- or marking errors roughly

But ESA is NEW.

ESA = **Humans mark the exact part of the translation where the error occurs.**

Example:

Source: "He didn't go."
Bad translation: "He go."

A human annotator marks the **exact place** where the mistake is:

```
He [go]  ← this span has an error (missing "didn't")
```

So ESA:

- highlights “error spans”
- shows type of error
- shows how serious each mistake is

This makes evaluation more accurate.

What is this paper about overall?

This paper is basically a **big report** summarizing:

✔ Which translation systems participated in WMT 2024
✔ What language pairs they worked on
✔ How well the translation systems performed
✔ How well LLMs performed
✔ How well online translators performed
✔ How the new ESA evaluation method helped judge results
✔ Which systems were the best and worst
✔ What trends they observed in translation quality

It’s a **complete overview** of the 2024 machine translation competition.

Main Concepts Used in the Abstract (Explained Simply)

1. Machine Translation (MT)

Automatically translating text from one language to another.

2. Language pair

Two languages used for translation (e.g., English ↔ German).

3. Domains

Different types of text used for testing, like:

- medical

- news
- blogs
- conversation

4. Large Language Models (LLMs)

Very big AI models like GPT and Llama, trained to understand and generate text.

5. Online translation providers

Translation websites like:

- Google Translate
- DeepL
- Microsoft Translator

6. Human annotators

People who read translations and judge whether they are correct.

7. ESA (Error Span Annotation)

A new evaluation method where human reviewers mark **exactly which words** in a translation are wrong.

SUPER SHORT SUMMARY IN 2 LINES

The paper gives the full results of the 2024 machine translation competition.
It compares translation systems, LLMs, and online translators using a new evaluation method called ESA.
