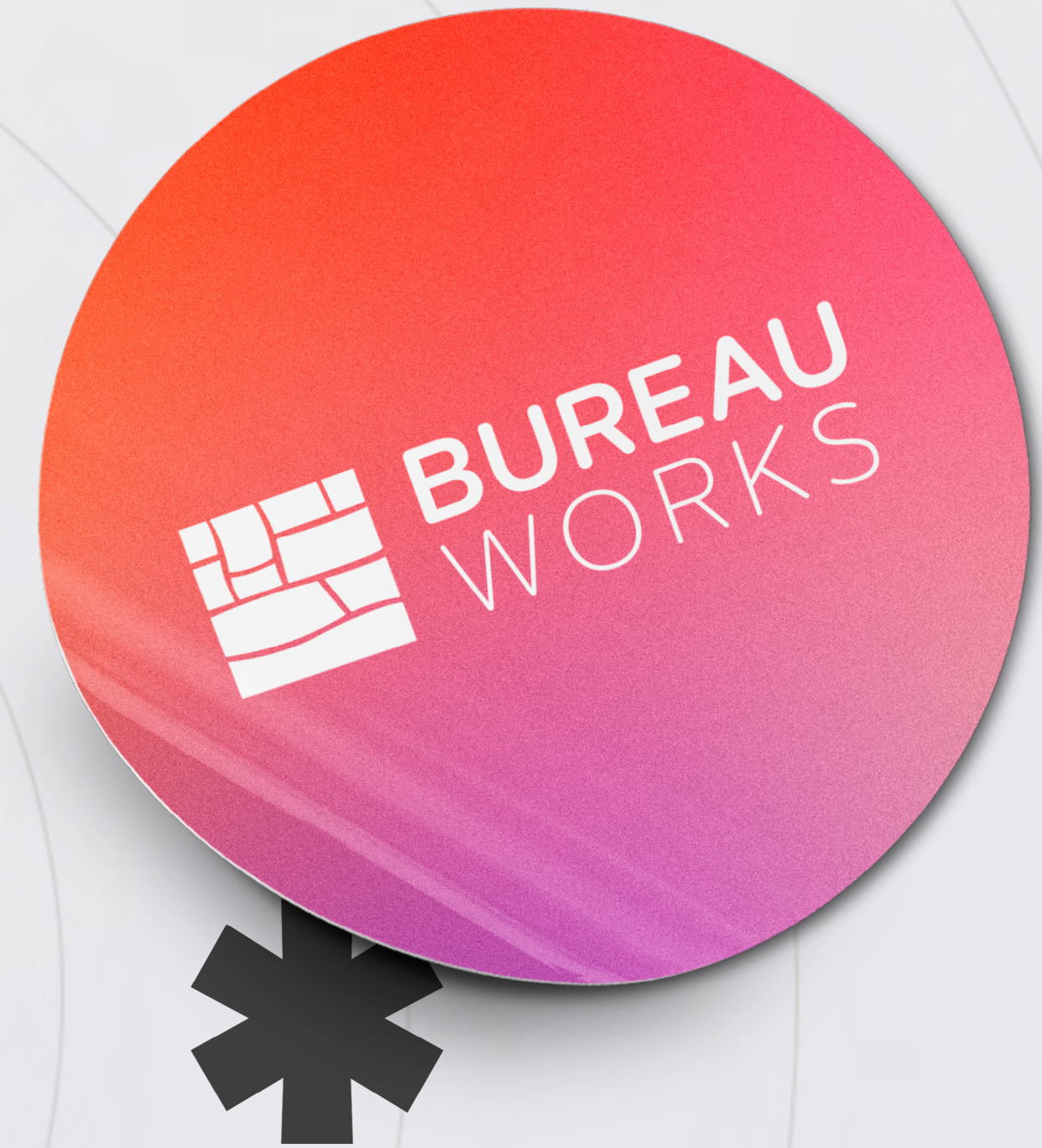


Chat GPT Vs. Amazon / Google / Microsoft

Machine Translation into Brazilian Portuguese,
Chinese, German, French, Italian,
Korean, Spanish

*A Study of Potential
Uses and Pitfalls using
Large Language Models
for Translation and
Linguistic Evaluation*



3	Abstract
4	Methodology
5	Disclaimers
7	Goals
8	Results per Language
9	Chinese
10	Output from a Qualitative Perspective
12	Evaluation of GPTs Evaluation
13	Analysis and Key Findings
15	Practical applications and limitations
16	French
17	Output from a Qualitative Perspective
19	Evaluation of GPTs Evaluation
20	Analysis and Key Findings
22	Practical applications and limitations
23	German
24	Output from a Qualitative Perspective
26	Evaluation of GPTs Evaluation
27	Analysis and Key Findings
28	Practical applications and limitations
30	Italian
31	Output from a Qualitative Perspective
33	Evaluation of GPTs Evaluation
34	Analysis and Key Findings
35	Practical applications and limitations
36	Korean
37	Output from a Qualitative Perspective
39	Evaluation of GPTs Evaluation
40	Analysis and Key Findings
41	Practical applications and limitations
42	Portuguese
43	Output from a Qualitative Perspective
45	Evaluation of GPTs Evaluation
46	Analysis and Key Findings
48	Practical applications and limitations
49	Spanish
50	Output from a Qualitative Perspective
52	Evaluation of GPTs Evaluation
53	Analysis and Key Findings
55	Practical applications and limitations
56	Conclusions



u * works * bureau * NO

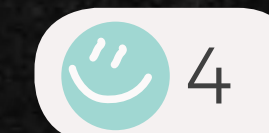
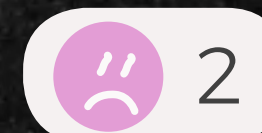
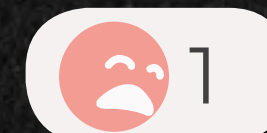
Abstract

Language is a tricky business. While everyone is blown away by the prowess and potential of language models such as GPT-3, we decided to take a **deeper look into the finer ability to understand, rate, and contextualize** metaphorical language and idiomatic expressions.

Methodology

We selected 10 common idiomatic expressions in English.

We provided them to Amazon, Google and Microsoft for non-tuned machine translation. We then asked GPT to evaluate these translations from 1 to 5, 5 relaying the right meaning in a natural way and 1 relaying the wrong meaning in an unnatural way.



We also asked GPT-3 to provide us with its translation and then a second iteration of its translation in a more figurative form.



Disclaimers

(Real research disclaimers not to be taken lightly)

While we aimed to explore the potential of ChatGPT's language capabilities, it is important to note that this study only evaluated one aspect of translation, namely the ability to handle linguistic idiomatic highly metaphorical edge cases. Other aspects of translation, such as cultural and contextual understanding, may require different evaluation methods and criteria.

The sample size of this study is limited to 10 idioms and one reviewer per language, which may not be representative of the full range of idiomatic expressions in the English language, or the range of perspectives and expertise of professional translators. As such, the results of this study should be interpreted with caution and cannot be generalized to other contexts or domains.

Furthermore, the opinions and evaluations of the single reviewer per language are subjective and may be influenced by personal biases, experiences, or preferences. As with any subjective evaluation, there is a degree of variability and uncertainty in the results. To increase the reliability and validity of our findings, future studies could involve multiple reviewers, blind evaluations, or inter-rater reliability measures.

It is also worth noting that ChatGPT's language capabilities are not static and may change over time as the model is further trained and fine-tuned. Therefore, the results of this study should be considered as a snapshot of the model's performance at a specific point in time, and may not reflect its current or future capabilities.

Lastly, this study is not intended to make any definitive or categorical claims about the usefulness or limitations of ChatGPT for translation. Rather, it is meant to serve as a preliminary investigation and starting point for future research and development in the field of natural language processing and machine translation. As with any emerging technology, there are still many challenges and opportunities for improvement, and further experimentation and collaboration will be needed to fully explore its potential.

Goals

The main objective of this study was to

1

Compare Machine Translation Outputs against each other

2

Establish patterns/ commonalities between Machine Translation outputs

3

Compare Machine Translation outputs against GPT-3

4

Compare GPT-3 first vs. second iteration

5

Evaluate GPT-3 ability to quantitatively and qualitatively assess translation quality

6

From these findings infer potential uses, limitations and pitfalls of language models

RESULTS

per

Language

Chinese

The following analysis was written by our English/Chinese linguistic James Hou.



Overall all engines struggled with the metaphorical nature of language, often erring in excessive literality. For example, The translation “让我们称之为一天。” is a more literal and somewhat awkward translation of “Let’s call it a day.” It conveys the meaning of the original phrase, but the wording is somewhat awkward and may not be as easy for native Chinese speakers to understand.

While it accurately conveys the meaning of the original phrase, the wording is somewhat awkward and may be less clear to native speakers.



















































Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into Chinese.

	Google	Amazon	Microsoft	GPT
Success Rate	40%	30%	60%	90%
Awkward Rate	30%	30%	20%	10%
I want to cry	30%	40%	20%	0%

As far as translation quality goes, GPT did a great job with contextualization. For example, “秘密被揭露了。” is an accurate and clear translation for “The cat’s out of the bag”. In 9 out of 10 sentences, the content was well adapted, intelligible, and conveyed the appropriate meaning. Contrary to the three Machine Translation engines, GPT had no embarrassing “I want to cry” mistakes.

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	猫从袋子里出来了。 	那只猫从袋子里拿出来了。 	猫从袋子里出来了。 	这只猫从袋子里出来了。 	秘密被揭露了。 	Most translations are more detailed or literal translations of "The cat's out of the bag", except the last two. "秘密被揭露了。" is an accurate and clear translation. ...See full review.
Let's call it a day.	让我们收工。 	让我们称之为一天。 	今天就到这儿吧。 	咱们今天就到这里吧。 	咱们今天的工作就到这里为止吧。 	Most translations are clear and accurate translations of "Let's call it a day." They convey the meaning of the original phrase well and are easy for native Chinese speakers to understand. ...See full review.
I'm head over heels for him.	我为他着迷。 	我对他不知所措。 	我为他着想。 	我对他一见钟情。 	我对他心动不已。 	Most translations are clear and accurate translations of "I'm head over heels for him." They convey the meaning of the original phrase well and are easy for native Chinese speakers to...See full review.
I get the hang of it.	我掌握了窍门。 	我明白了。 	我掌握了窍门。 	我懂了。 	我已经掌握了。 	All translations are clear and accurate translations of "I get the hang of it." They convey the meaning of the original phrase well and are easy for native Chinese speakers to understand.
Hit me up.	打我。 	打我。 	打我。 	给我打个电话。 	给我发个信息。 	Some translations are "打我。", not accurate or appropriate translation of "Hit me up." The last three translations are accurate or appropriate translations.
It really took the wind out of my sails.	它真的让我失去了风帆。 	它真的挡住了我的风帆。 	它真的把风从我的帆上带走了。 	这真的让我失落沮丧。 	这真的让我打了个懵。 	Some translations are not accurate or appropriate translations. The last three translations are accurate or appropriate translations.
They are two peas in a pod.	它们是豆荚里的两颗豌豆。 	它们是豆荚里的两颗豌豆。 	它们是豆荚里的两颗豌豆。 	他们俩就像一对好搭档。 	他们俩就像一心一意。 	All translations are less accurate translations of "They are two peas in a pod."
It takes two to tango.	一个巴掌拍不响。 	探戈需要两个人的时间 	一个巴掌拍不响。 	这是个两个人的事。 	这是个需要两个人配合的事。 	The translation "一个巴掌拍不响。" is a less accurate translation of "It takes two to tango." Other translations are clear and accurate translations,conveying the meaning of the original...See full review.
It's a piece of cake.	很简单的。 	这是小菜一碟。 	这是小菜一碟。 	这很简单。 	这是小菜一碟。 	All translations are clear and accurate translations of "It's a piece of cake." They convey the meaning of the original phrase well and are easy for native Chinese speakers to understand, ...See full review.
That costs an arm and a leg.	那要花一条胳膊和一条腿。 	这需要一只手臂和一条腿。 	这要花一条胳膊和一条腿。 	那很贵。 	那很破费。 	The translation "这需要一只手臂和一条腿。" is a less accurate translation of "That costs an arm and a leg." ...See full review.

Evaluation of GPTs Evaluation

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	70%	70%	50%
Quantitative Analysis Accuracy	60%	60%	40%

Analysis and Key Findings

- Google and Amazon had extremely similar results, only slightly deviating from each other, mirroring each other's mistakes and metaphorical choices. For example, the translation “打我。” is not an accurate or appropriate translation of “Hit me up.” The phrase “Hit me up” means to contact or get in touch with someone, typically by phone or text message. The Chinese phrase “打我” means “hit me,” and it does not convey the meaning of the original phrase.

- Microsoft made bolder choices when it comes to the linguistic adaptation of the idioms.

- GPT had a harder time evaluating Microsoft's metaphorical choices as they departed more.

- GPT-3 had an easier time with the Qualitative Analysis producing cogent textual analysis (even though with only 70% accuracy).

- Although intelligible GPT's analysis failed to identify in 30% of the cases. This coincided with metaphorical choices that were literal and understandable but deviated from quotidian discourse.

- GPT-3 had a harder time translating the qualitative analysis into a score. Although broadly speaking scores were 60% accurate, it was difficult to differentiate between similar scores such as a 3 vs. a 4.

- Extreme score divergence from 1 to 5 was easier to understand and more compatible with overall comments suggesting that:
 - Perhaps scoring criteria was not sufficiently calibrated with GPT-3
 - Perhaps binary scoring could be more relevant than gradient scoring

- Even though quantitatively Microsoft performed similarly to Google and Amazon, when you get into the nitty gritty of language, Microsoft made bolder choices and provided better results from a qualitative perspective but still was far behind GPT-3 when it came to accuracy and cultural adaptation. For example, The translation “我掌握了窍门。” is a clear and accurate translation of “I get the hang of it.” It conveys the meaning of the original phrase well and is easy for native Chinese speakers to understand. It is also a more figurative and idiomatic way of expressing the idea of understanding or mastering something compared to other possible translations.

Practical applications and limitations

In this analysis GPT-3 provided superior contextualization and adaptation than previous machine translation models.

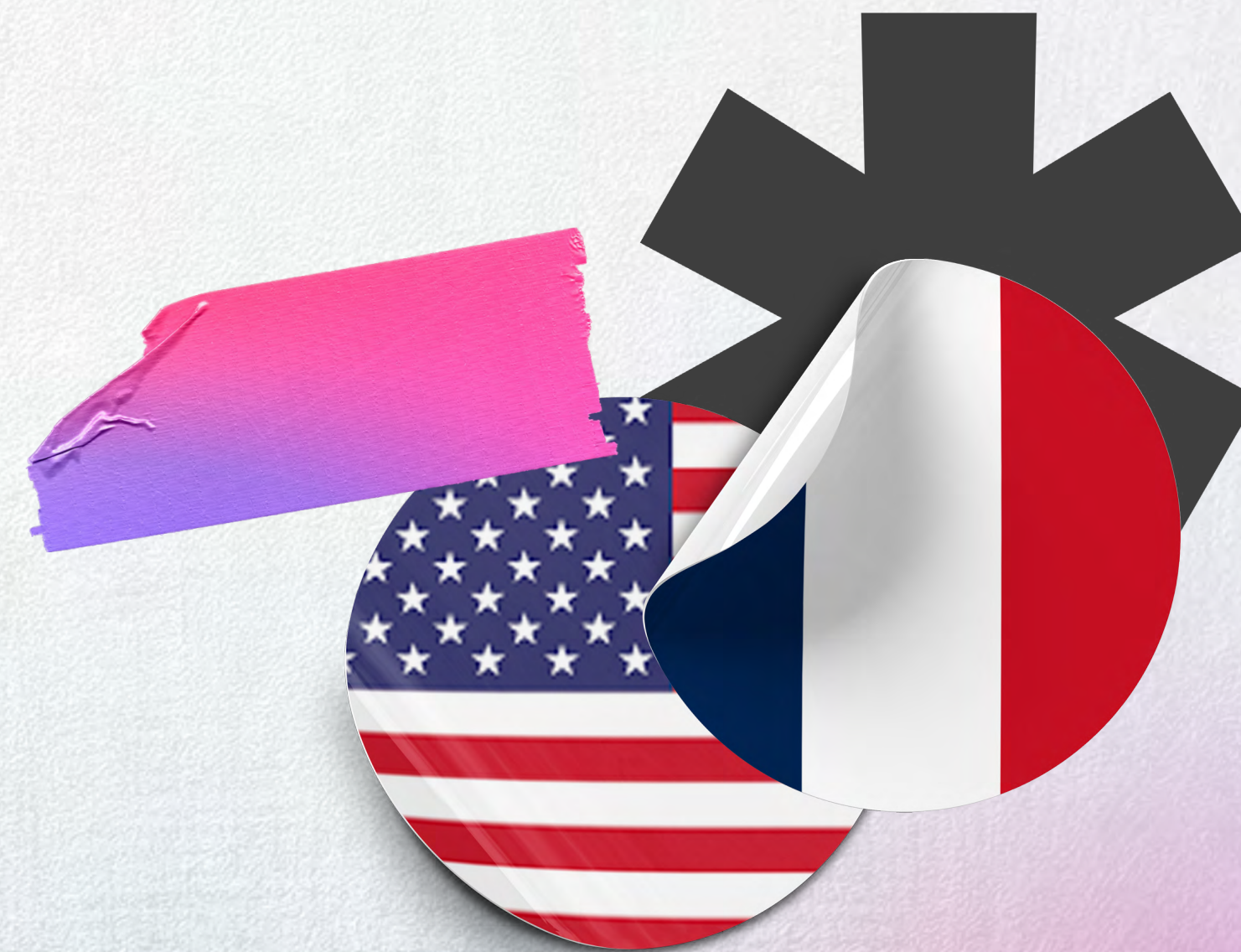
While none of the engines are reliable enough to replace humans (at least in the context of this study), GPT-3 shows clear capability of aiding human translators and reviewers in the process of translating and evaluating language.

French

The following analysis was written by our English/French linguistic Laurène Bérard.



Overall all engines struggled with the metaphorical nature of language, often erring in excessive literality.



Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into French.

	Google	Amazon	Microsoft	GPT
Success Rate	20%	30%	40%	80%
Awkward Rate	20%	20%	10%	10%
I want to cry	60%	50%	50%	10%

As far as translation quality goes, GPT did a great job with contextualization. Without any guidance, GPT was totally wrong for “The cat’s out of the bag” but was right with more guidance, and its 3 translations were grammatically incomplete thus quite difficult to understand for “It takes two to tango”.

The content was generally well adapted, intelligible, and conveyed the appropriate meaning. The three Machine Translation engines had a high “I want to cry” rate while even without context a translator would have guessed it wasn’t meant literally (e.g. “Let’s call it a day”, which is quite obvious but was totally misunderstood by Google engine).

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	Le chat est sorti du sac. 😞	Le chat est sorti du sac. 😞	Le chat est sorti du sac. 😞	Le chat est sorti du sac. 😞	Le secret est évené. 😊	All three Machine Translation outputs were literal translations which did not accurately relay meaning. GPT's original translation was also literal but when prompted to be more figurative...See full review.
Let's call it a day.	Appelons le un jour. 😞	Disons que c'est fini. 😊	On s'arrête là pour aujourd'hui. 😞	Terminons là pour aujourd'hui. 😊	Mettons un terme à cette journée. 😞	French Google is the worst translation, completely literal and not understandable by French people (and it includes a spelling mistake, as it should be "Appelons-le". GPT was right in its comments for each, but a score of 3 ...See full review.
I'm head over heels for him.	Je suis folle de lui. 😊	Je suis éperdument amoureux de lui. 😊	Je suis folle pour lui. 😞	Je suis follement amoureux de lui. 😊	Je suis raide dingue de lui. 😊	GPT rating for French Microsoft is wrong, as "'Je suis folle pour lui'" has no meaning in French. We would guess what the person is trying to say if they are foreign people, ...See full review.
I get the hang of it.	Je comprends. 😊	Je m'y habitue. 😊	Je comprends bien. 😊	Je commence à comprendre. 😊	Je commence à voir comment ça marche. 😊	GPT rating for French Amazon and Google is not completely right as the translation heavily depends on context here. If the context is right, Google and Amazon translations may ...See full review.
Hit me up.	Battez moi. 😊	Frappe-moi. 😞	Frappez-moi. 😞	Contacte-moi. 😊	Appelle-moi. 😊	Again, Google translation is totally wrong and literal and includes again a typo as it should be written "'Battez-moi'". Amazon and Microsoft are totally wrong too, but at least ...See full review.
It really took the wind out of my sails.	Cela a vraiment coupé le vent de mes voiles. 😞	Cela m'a vraiment fait perdre le vent. 😞	Cela m'a vraiment coupé l'herbe sous le pied. 😊	Cela m'a vraiment découragé. 😊	Cela m'a coupé l'herbe sous le pied. 😊	GPT assessment for Google is wrong. Google translation is purely literal and won't ring a bell to any French speaker. Both Google and Amazon ...See full review.
They are two peas in a pod.	Ce sont deux pois dans une cosse. 😞	Ce sont deux petits pois dans une gousse. 😞	Ce sont deux pois dans une gousse. 😞	Ils sont comme les deux doigts de la main. 😊	Ils sont inséparables. 😊	Here, GPT assesement and rating for all 3 translations is totally wrong. These translations are literal, are not commonly used translations in French for this saying and cannot be understood as such in French....See full review.
It takes two to tango.	Il faut être deux pour danser le tango. 😊	Il faut être deux pour danser le tango. 😊	Il faut être deux pour valser. 😊	Il faut deux pour tango. 😞	Il faut deux pour faire la paire. 😞	Amazon and Google translations are fine, as the English can be similar to the French here, although we could translate it without the tango reference...See full review.
It's a piece of cake.	C'est un morceau de gâteau. 😞	C'est un jeu d'enfant. 😊	C'est du gâteau. 😊	C'est un jeu d'enfant. 😊	C'est un jeu d'enfant. 😊	Unlike what GPT assessed, Google translation is wrong. It is a literal translation that would only work if we are actually talking about the piece of a cake....See full review.
That costs an arm and a leg.	Cela coûte un bras et une jambe. 😞	Cela coûte un bras et une jambe. 😞	Cela coûte un bras et une jambe. 😞	Cela coûte les yeux de la tête. 😊	Cela coûte une fortune. 😊	"Again, GPT is wrong when assessing the three providers. These are literal translations that would deserve a score of 2 or maximum 3 as they are not used in French. We would probably ...See full review.

Evaluation of GPTs Evaluation

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	50%	70%	50%
Quantitative Analysis Accuracy	50%	70%	50%

Analysis and Key Findings

- Despite translating correctly 8 times out of 10, GPT analysis of the 3 engines was often wrong, as it almost 50% of the cases stated that their translation was fine while it was not.
- Although intelligible GPT's analysis failed to identify issues in 30 to 50% of the cases, this coincided with choices that were literal and would have been correct if not dealing with idiomatic expressions.

- GPT failed to spot the two spelling issues from Google ("Appelons le" and "Battez moi").
- GPT qualitative and quantitative analyses were globally consistent with each other.
- GPT had a harder time evaluating Microsoft's and Google's translations.
- Microsoft made better choices when it comes to the linguistic adaptation of the idioms.

french * fr
*
*
youna

- Google and Amazon had very similar results, only slightly deviating from each other. Microsoft stood out from the two.

- In one case (the tango expression), GPT evaluated a fine translation from Microsoft as being wrong (score of 2) while it was understandable and equally good as Google and Amazon translations which GPT evaluated positively (score of 5), but it may be explained by the fact that Microsoft was not as literal as the other two engines, as it chose to mention walz instead of tango.

Microsoft made bolder choices and provided better results from a qualitative perspective but still was far behind GPT when it came to accuracy and cultural adaptation.

Practical applications and limitations

In this analysis GPT provided far better contextualization and adaptation than Amazon, Google and Microsoft machine translation engines. While it's not convenient to replace traditional Machine Translation models with larger models such as GPT-3 due to high computational costs and diminishing marginal gains when it comes to non-metaphorical discourse, GPT-3 can be a powerful human ally when it comes to providing suggestions and identifying potential mistakes as well as opportunities for improvement.

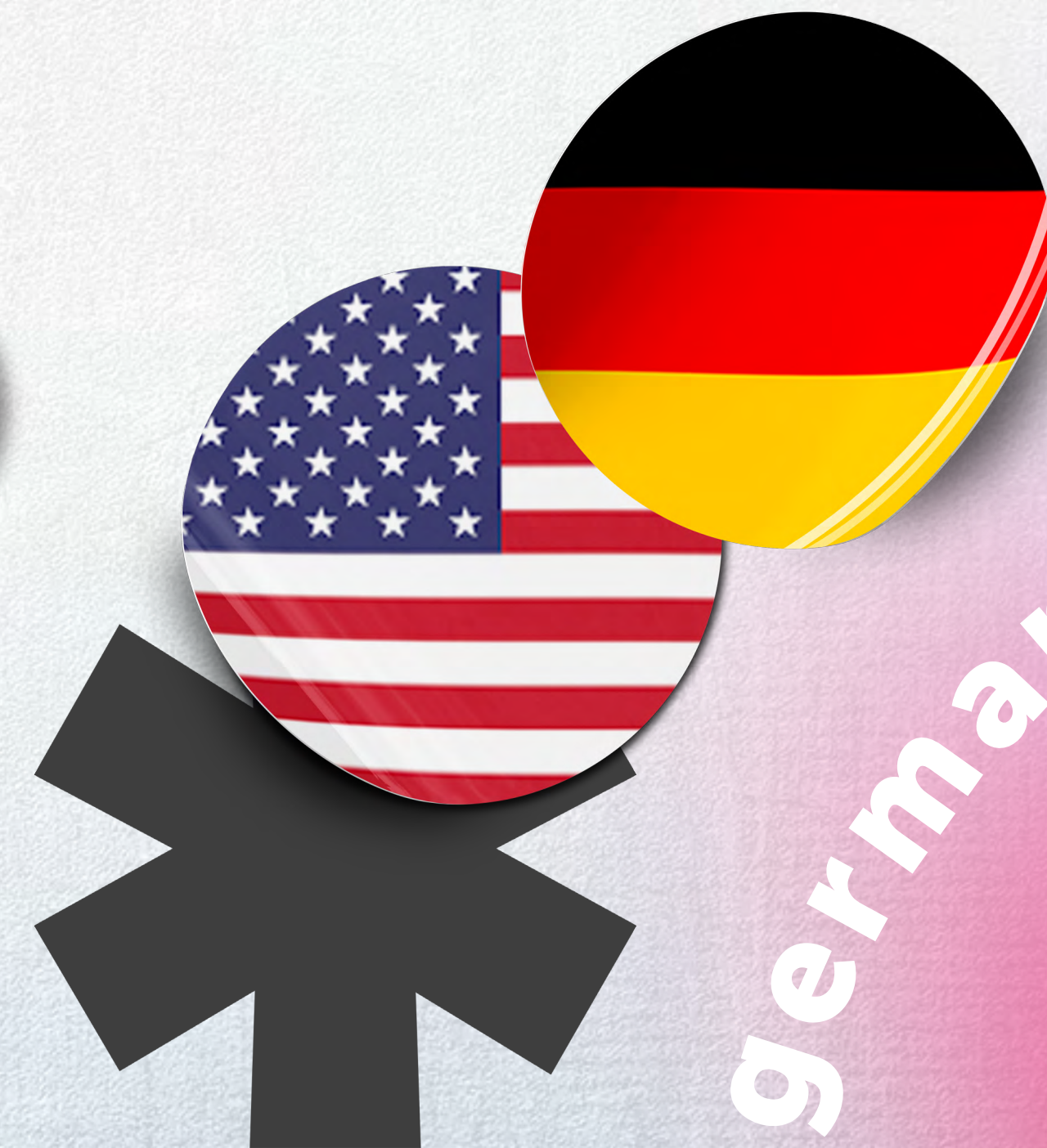
Even though machine translation engines are “nearly” there, this “nearly” becomes progressively harder to tackle as proven by GPT here.

While none of the engines are reliable enough to replace humans (at least in the context of this study), GPT shows clear capability of aiding human translators and reviewers in the process of translating and evaluating language.

German

The following analysis was written by our English/German linguistic Olga Schneider.

Overall, GPT produced the most accurate translation. It always analyzed the English sentence accurately and could usually tell whether a machine translation was literal or idiomatic, but it failed to detect a mistranslation 50% of the time.



* german *
* german *
* german *

Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into German.

	Google	Amazon	Microsoft	GPT
Success Rate	30%	50%	50%	80%
Awkward Rate	20%	10%	10%	0%
I want to cry	50%	40%	40%	20%

GPT produces good results 8 out of 10 times on the first try and 7 out of 10 times on the second try. For example, only GPT was able to provide accurate idiomatic translations for “I’m head over heels for him” and “Hit me up”. The last, most figurative output produced 4 good sentences, with the others being either all right or “I want to cry”. Overall, Amazon and Microsoft produced slightly better translations than Google.

*french * german * ge

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	Die Katze ist aus dem Sack. 😊	Die Katze ist aus dem Sack. 😊	Die Katze ist aus dem Sack. 😊	Die Katze ist aus dem Sack. 😊	Das Geheimnis ist gelüftet. 😞	All three MTs produces the same sentence, which is a literal translation and perfectly appropriate because we use the same expression in German. Chat GPT....See full review.
Let's call it a day.	Machen wir für heute Schluss. 😊	Machen wir Feierabend. 😊	Jetzt wird Feierabend gemacht. 😞	Lass uns für heute Schluss machen. 😊	Lass uns den Vorhang fallen lassen. 😊	All MT translations were correct and GPT accurately evaluated them. The one from Microsoft is correct but it can't be used in every context, it will highly depend on the tone you want...See full review.
I'm head over heels for him.	Ich bin Hals über Kopf für ihn. 😞	Ich bin Hals über Kopf für ihn. 😞	Ich bin Hals über Kopf für ihn. 😞	Ich bin hoffnungslos in ihn verliebt. 😊	Ich bin bis über beide Ohren in ihn verliebt. 😞	Only ChatGPT provided an accurate translation. Its ratings were 5, although it rather should have been 3. The other MTs provided an almost correct translation, but it sounds incomplete...See full review.
I get the hang of it.	Ich verstehe es. 😞	Ich hab den Dreh raus. 😊	Ich habe den Dreh raus. 😊	Ich kapiere es. 😞	Ich habe den Dreh raus. 😞	All translations were correct and could be used this way in a text. GPT accurately rated Google with a 3, as it conveyed the meaning but it didn't use an idiomatic German expression...See full review.
Hit me up.	Schlag mich an. 😞	Schlag mich an. 😞	Schlag mich. 😞	Schreib mir. 😞	Gib mir Bescheid. 😞	GPT accurately rated the other MTs with 1 and gave an accurate explanation for its evaluation. Its own translations were correct and could be used in a text.
It really took the wind out of my sails.	Es hat mir wirklich den Wind aus den Segeln genommen. 😞	Es hat mir wirklich den Wind aus den Segeln genommen. 😞	Es hat mir wirklich den Wind aus den Segeln genommen. 😞	Es hat mir den Wind aus den Segeln genommen. 😞	Es hat mir den Boden unter den Füßen weggezogen. 😞	All MTs provided a literal translation of the expression because in German, we have an expression that is a literal translation of this phrase but with a...See full review.
They are two peas in a pod.	Sie sind zwei Erbsen in einer Schote. 😞	Sie sind zwei Erbsen in einer Schote. 😞	Es sind zwei Erbsen in einer Schote. 😞	Sie sind zwei Herz und eine Seele. 😞	Sie sind wie Siamesische Zwillinge. 😊	The English expression is "used to say that two people or things are very similar to each other". GPT accurately explained the English meaning. However, it rated....See full review.
It takes two to tango.	Es gehören immer zwei dazu. 😊	Zum Tango braucht man zwei. 😞	Es gehören zwei dazu. 😊	Es braucht zwei, um Tango zu tanzen. 😞	Es braucht zwei, um einen Streit zu schlichten. 😊	GPT accurately explained the meaning of the English expression but it failed in its evaluation of the MT. While Google and Microsoft were accurate...See full review.
It's a piece of cake.	Es ist ein Stück Kuchen. 😞	Es ist ein Kinderspiel. 😊	Das ist ein Kinderspiel. 😊	Das ist ein Kinderspiel. 😊	Das ist ein Klacks. 😞	GPT accurately explained the meaning of the English phrase and also accurately rated the MTs, giving Google the rating of 2 because it missed the point (should have been 1 actually)...See full review.
That costs an arm and a leg.	Das kostet einen Arm und ein Bein. 😞	Das kostet einen Arm und ein Bein. 😞	Das kostet einen Arm und ein Bein. 😞	Das kostet ein Vermögen. 😊	Das kostet ein Heidengeld. 😞	GPT accurately understood the English meaning and provided two perfect suggestions. It rated the MT with 5 but even though this expression will be understood easily ...See full review.

Evaluation of GPTs Evaluation

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	50%	50%	50%
Quantitative Analysis Accuracy	50%	50%	50%

Analysis and Key Findings

- GPT has a fairly good knowledge of German idiomatic expressions, even the ones that did not quite match the English sentence were idiomatic expressions of a similar category. For example: “Let’s pull the plug” and “Let’s drop the curtain” for “Let’s call it a day”. It understands the general meaning of these phrases related to “ending” and could be a useful source of inspiration.
- When it comes to creative expressions, GPT seems to take inspiration from English phrases, suggesting the literal German translation of “I am in love with him from head to toe” as a creative alternative to “I’m head over heels for him”.
- GPTs translation ratings were hit or miss and proved to be unreliable. In one case, it rated a Google translation a 2, and it was not clear why it was not the 1 it should have been.
- German and English share many idiomatic expressions, which makes translation easier. But expressions that are foreign to German (e.g. “They are two peas in a pod” or “It’s a piece of cake”) end up being translated literally. In the case of the peas, unlike the translation engines, GPT understood that it needed to provide expressions about “twos”. However, the expressions it provided – while accurate and commonly used – didn’t convey the correct meaning.

Practical applications and limitations

GPT provided better translations than Google, Amazon and Microsoft.

While it is not 100% reliable, it can provide a better starting point for machine translation editing than the other three engines. While idiomatic expressions are important, there is another problem often encountered in machine translation of English to German text: cumbersome sentence structure that is too close to the original text. It would be important to see how GPT solves this problem.

When it comes to evaluation, GPT is not a good tool, as its evaluation of German translations is only 50% accurate.

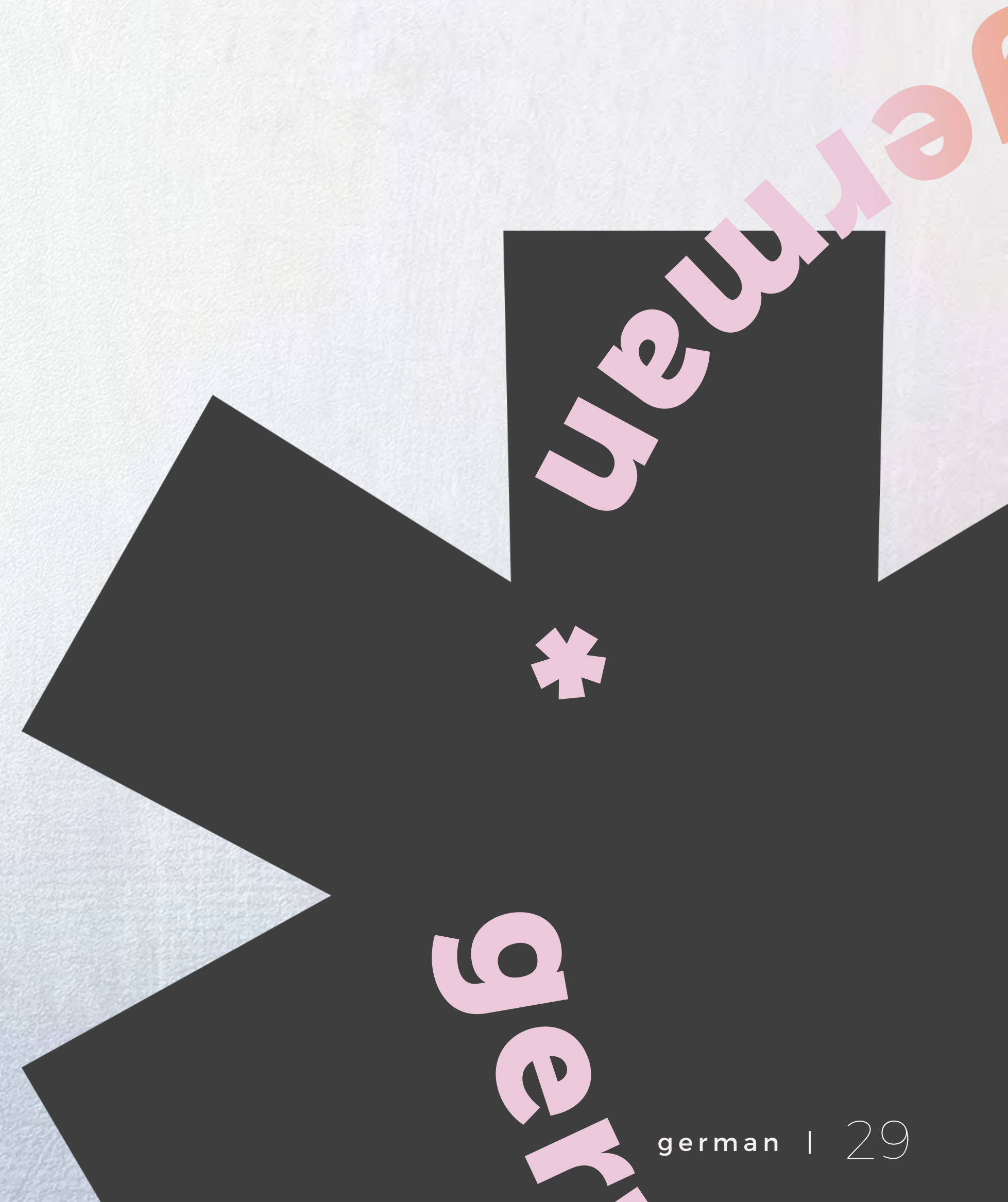
Three more things would have been interesting to see:

1) Can GPT compete with DeepL for German? While GPT may provide good translations, DeepL produces good German translations and also offers a range of features that simplify the translation process (glossary terms translated correctly with the correct plural and case, one-click editing, autocomplete sentences after typing one or two words to speed up rephrasing). GPT's translation needs to be significantly better than that of DeepL to make up for the lack of features.

2) Could GPT's accuracy be improved with more context, such as a paragraph containing the phrase?

3) If Google is able to recognize AI-generated text, how will it handle GPT translations with minimal to no editing? Can it detect the "GPT style" and penalize a text in its search results?

In summary, in its current state, GPT can be a source of inspiration for our limited human brains. It can provide decent translations. More than that, it can help us rephrase overused expressions, find metaphors, and think outside the box.

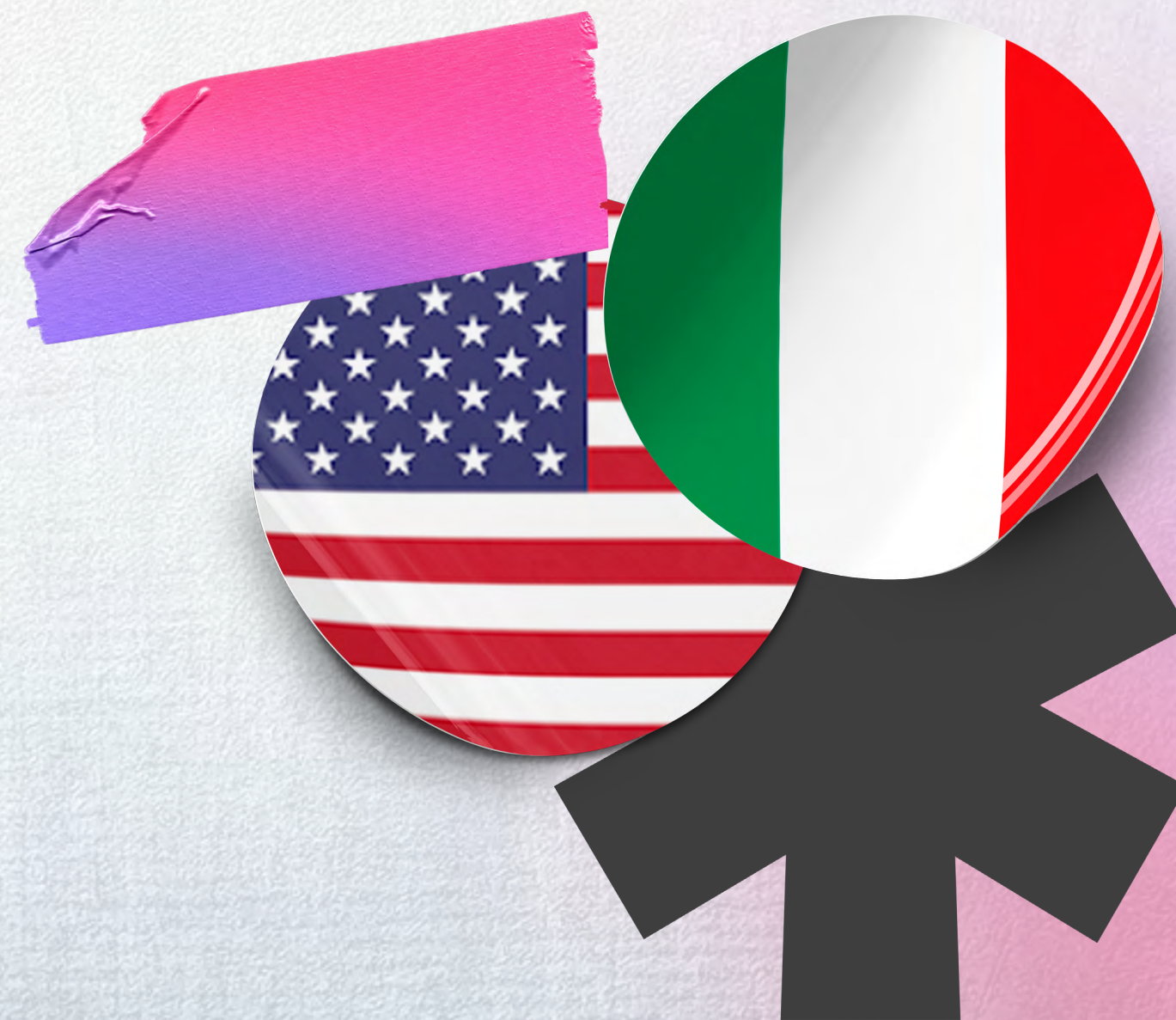


Italian

The following analysis was written by our English/Italian linguistic Elvira Bianco.



Overall all engines struggled with the metaphorical nature of language, often erring in excessive literality.



Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into Italian.

	Google	Amazon	Microsoft	GPT
Success Rate	34%	36%	35%	76%
Awkward Rate	66%	64%	65%	24%
I want to cry	60%	40%	50%	10%

Machine translation gave a literal translation very far from the correct meaning. GPT conveyed the correct meaning and expression at 70%, for the remaining 30% used acceptable expressions that are not widely used or perceived like natural native speech.

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	Il gatto è fuori dal sacco. 😞	Il gatto è uscito dalla borsa. 😞	Il gatto è fuori dal sacco. 😞	Ecco come stanno le cose. 😊	La gatta è uscita dal sacco. 😞	Before evaluating these translations I need to point out that when we translate proverbs, sayings and maxims we should look for the right expression that conveys the same or a ...See full review.
Let's call it a day.	Chiamiamolo un giorno. 😞	Chiamiamola giornata. 😊	Chiudiamola qui. 😊	Facciamo finire qui la giornata. 😊	Facciamo una pausa. 😊	Same as above, when we translate a saying we can't use any literal translation, but convey the right meaning we another saying commonly used in target language...See full review.
I'm head over heels for him.	Sono perdutoamente innamorato di lui. 😞	Sono perdutoamente innamorato di lui. 😞	Sono testa sopra i tacchi per lui. 😞	Sono pazzamente innamorato di lui. 😞	Sono pazzamente innamorato di lui. 😞	I would give score 4 to Google and amazon, both translations give the meaning but look like a man is in love with another man (not woman in love give a man); Microsoft translation...See full review.
I get the hang of it.	Ho capito. 😊	Ho capito come funziona. 😊	Ho capito. 😊	Ci sono arrivato. 😊	Ho capito come funziona. 😊	All translations are correct in meaning and grammar. I would give 4 score to Google and Microsof, 5 score to Amazon because it is more exhaustive. GPT1 = 4 because it may not fit ...See full review.
Hit me up.	Mi ha colpito. 😞	Mi chiami. 😊	Colpiscimi. 😞	Contattami. 😊	Mandami un messaggio. 😊	Googe translate here gets 1 score, it's totally wrong. Amazon gets 4 score (close to meaning) Amazon gets 1 score. GPT -1 is correct (5 score) GPT-2 (4 score) it's not sure the way we choose to contact.
It really took the wind out of my sails.	Mi ha davvero tolto il vento dalle vele. 😞	Mi ha davvero tolto il vento dalle vele. 😞	Mi ha davvero tolto il vento dalle vele. 😞	Mi ha veramente demoralizzato. 😞	Mi ha veramente sconvolto. 😞	Google, Amazon and Microsoft get 1 score, translation is literal and doesn't give the meaning. GPT-1 gets 4, it's close to meaning, GPT-2 gets 3 score, not the right meaning. Correct ...See full review.
They are two peas in a pod.	Sono due piselli in un baccello. 😞	Sono due piselli in un baccello. 😞	Sono due piselli in un baccello. 😞	Sono due gocce d'acqua. 😊	Sono come due gocce d'acqua. 😊	All machine translations get 1 score since literal and wrong in meaning. All GPT gets 5 (right meaning and grammar).
It takes two to tango.	Bisogna essere in due per ballare il tango. 😞	Bisogna essere in due per ballare il tango. 😞	La colpa non è mai di una sola persona. 😞	Ci vogliono due persone per fare il tango. 😞	Ci vogliono due persone per creare una situazione di conflitto. 😞	All machine translations are fine and give the correct meaning, so I would give 4 score to machine translation. GPT 1 is wrong, "tango" can be danced not done, score would ...See full review.
It's a piece of cake.	È un pezzo di torta. 😞	È un gioco da ragazzi. 😊	È un gioco da ragazzi. 😊	È facile come bere unbicchiere d'acqua. 😊	È semplice come mangiare un panino. 😞	Google translate gets 1 score, Amazon and Microsoft get 5, translation is correct in grammar and meaning. GPT -1 gets 5 score, GPT-2 gets 4 score, gets the meaning but ...See full review.
That costs an arm and a leg.	Costa un braccio e una gamba. 😞	Costa un braccio e una gamba. 😞	Questo costa un braccio e una gamba. 😞	Costa un occhio della testa. 😊	Costa una fortuna. 😊	All machine translations are literal and don't give the meaning, score is 1 for all machine translations. GPT-1 and GPT-2 get 5 score since these translations are both correct and widely used.

Evaluation of GPTs Evaluation

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	30%	40%	50%
Quantitative Analysis Accuracy	30%	40%	50%

Analysis and Key Findings

- Google Translate get only one right translation and get close to meaning in 2 sentences.
- Amazon gave 50% right meaning even if not using the most common way to convey the English saying in Italian language.
- Microsoft gave 3 right answer getting closer to Italian similar sayings.
- While 1st Chat GPT usually approves machine translations, 2nd and 3rd Chat GPT usually give the correct meaning and add valuable translation suggestions.



Practical applications and limitations

As defined by https://it.wiktionary.org/wiki/espressione_idiomatica an idiomatic expression typical of a language is usually untranslatable literally into other languages except by resorting to idiomatic expressions of the language into which it is translated with meanings similar to the idiomatic expressions of the language from which it is translated. Clearly the mechanical translation

produced by today's most used translation machines (Google, Amazon, Microsoft) was unreliable, notwithstanding GPT-3 proved to be able at 70% to give the right meaning and to furnish good suggestions in content adaptation.

It is not unlikely that in the near future machines will memorize also idiomatic expressions but right now we need humans to translate conveying the same meaning from one language to another.

Languages are full of nuances, double entenders, allusions, idioms, metaphors that only that only a human can perceive.

italian

Korean

The following analysis was written by our English/Korean linguistic Sun Min Kim.



Overall all engines struggled with the metaphorical nature of language, often erring in excessive literality.



Output from a Qualitative Perspective



















































The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into Korean.

	Google	Amazon	Microsoft	GPT
Success Rate	0%	10%	30%	40%
Awkward Rate	40%	20%	20%	30%
I want to cry	60%	70%	50%	30%

Most engines literally translated the idiomatic expressions, while GPT tries to translate as descriptive as possible using no metaphor (e.g. piece of cake = easy, while in Korea, we have a similar idiomatic expression that conveys the same meaning as Microsoft did.

GPT's three translations are not consistent. Some are getting worse with the iteration.

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	고양이가 가방에서 나왔습니다. 	고양이가 가방 밖으로 나왔어요. 	고양이가 가방에서 나왔습니다. 	고양이가 주머니에서 나왔어요. 	고양이가 상자에서 나왔어요. 	3rd Chat GPT may be thought of most correct. However, it is a literal translation of "a secret has been revealed." Because we don't have a corresponding idiomatic expression in...See full review.
Let's call it a day.	오늘은 끝내자. 	이걸 하루라고 부르죠. 	오늘은 여기까지하자. 	그만 그만 하자. 	그만해요. 	The best is Microsoft. Google and 2nd Chat GPT translations can have other meanings, so I don't think they are good. Amazon and Chat GPT Original are not acceptable. 3rd Chat GPT went to far.
I'm head over heels for him.	나는 그를 위해 머리를 숙이고 있습니다. 	난 그를 정말 좋아해. 	나는 그를 위해 발 뒤꿈치를 넘고 있습니다. 	그 사람을 정말 사랑해요. 	그 사람에게 정신을 잃어버렸어요. 	Amazon and all three GPT are good. Others are not acceptable.
I get the hang of it.	요령을 이해합니다. 	요령은 알겠어. 	나는 그것의 요령을 얻는다. 	그걸 이제 이해했어요. 	그걸 이해했어요. 	Microsoft and 3rd GTP are wrong. Others convey the meaning correctly but are not naturally read. Maximum 3 points for all of them.
Hit me up.	띄워 줘. 	때려 줘. 	날 때려. 	연락해 주세요. 	저에게 연락해줘. 	Only Chat GPT's three translations convey the correct meaning. Others are not acceptable. 2nd GTP is wrong from the formal-informal consistency.
It really took the wind out of my sails.	정말 내 돛에서 바람이 불었습니다. 	정말 제 돛에서 바람을 막아줬어요. 	그것은 정말로 내 돛에서 바람을 제거했습니다. 	그거 정말 저의 힘을 차가웠어요. 	그게 정말 저의 희망을 꺾었어요. 	Other than 2nd GPT, everything is wrong. Again, I persoanlly think because there's no corresponding idiomatic expression in Korea, the literal translation of the metaphor can be accepted.
They are two peas in a pod.	꼬투리에 있는 두 개의 완두콩입니다. 	꼬투리에 두 개의 완두콩이 들어 있습니다. 	그들은 꼬투리에 두 개의 완두콩입니다. 	그들은 쌍둥이 같아요. 	그들은 같은 사람이에요. 	Other than GPT, all are literal. Again, I persoanlly think because there's no corresponding idiomatic expression in Korea, the literal translation can be accepted.
It takes two to tango.	탱고에는 두 명이 필요합니다. 	탱고에는 두 명이 필요합니다. 	손바닥도 마주쳐야 소리가 난다. 	탱고 춤은 두 사람이 해야 해요. 	그건 두 사람이 필요한 일이에요. 	Microsoft translation is the corresponding idiomatic expression in Korea. So it would be the best translation. Literal translation of the English can't be accepted because ...See full review.
It's a piece of cake.	그것은 케이크 한 조각입니다. 	케이크 한 조각이에요. 	식은 죽 먹기다. 	그건 쉬운 일이에요. 	그게 쉬운 일이에요. 	Microsoft translation is the corresponding idiomatic expression widely used in Korean. The three GPT Chat translations are descriptive (It is easy) and doesn't convey the feelings of the metaphor. Others are all literal and has no meaning in Korean.
That costs an arm and a leg.	그것은 팔과 다리 비용입니다. 	팔과 다리가 필요합니다. 	그것은 팔과 다리가 필요합니다. 	그거 정말 그만쳐요. 	그건 팔과 다리를 잃게 비싸요. 	Only the 3rd GPT is correct but it also went too far. It's funny because when it evaluated other translations it gave the most appropriate translations...See full review.

Evaluation of GPTs Evaluation

GPT knows the problems when the translations go wrong. But it is not considered to be correct in judging the best one.

GPT itself has issues with the consistency in sentence (e.g. formal – informal treatment and so on) and it can evaluate this issue.

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	60%	60%	60%
Quantitative Analysis Accuracy	30%	70%	50%

Analysis and Key Findings

- Because the English originals in this study are idiomatic expressions, it is a little bit tricky because you have to choose between the metaphor or direct description. But for some idiomatic expressions where the Korean language has similar idiomatic expression conveying the same meaning, most of the engines missed to find those expressions with only few exceptions (please see the worksheet and find those with the score of 5 by me).

- For others, my personal thought is that if the metaphor itself can convey the meaning, literal translation of it can be considered and maybe better to use a descriptive word. Of course, if the metaphor has no cultural context in Korea, it should not be literally translated. But it's a subtle issue and maybe up to the preference or human emotions of the translator. I don't think any engine has that level of human-like thinking yet.

Practical applications and limitations

I do think most of the engines can be used for pre-translation purpose. But considering the quality, it should be primarily for the efficiency purpose only (that is, not typing from the scratch). For more descriptive texts, such as manual, I see MTPE is much more advanced than these idiomatic expressions. So, there are still room to improve.

Portuguese

The following analysis
was written by our English/
Portuguese linguistic
Gabriel Fairman.



Overall all engines
struggled with the
metaphorical nature
of language, often erring
in excessive literality.



Portuguese * portu

* Portuguese * portu

Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into Portuguese.

	Google	Amazon	Microsoft	GPT
Success Rate	40%	40%	40%	90%
Awkward Rate	50%	50%	50%	10%
I want to cry	10%	10%	10%	0%

Most engines literally translated the idiomatic expressions, while GPT tries to translate as descriptive as possible using no metaphor (e.g. piece of cake = easy, while in Korea, we have a similar idiomatic expression that conveys the same meaning as Microsoft did.

GPT’s three translations are not consistent. Some are getting worse with the iteration.

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	O gato está fora do saco. 😞	O gato está fora da sacola. 😞	O gato está fora do saco. 😞	"O gato está fora da bolsa" 😞	"o segredo está no vento" 😊	All three Machine Translation outputs were literal translations which did not accurately relay meaning. See full review.
Let's call it a day.	Vamos encerrar o dia. 😊	Vamos encerrar o dia. 😊	Chega por hoje. 😊	Vamos encerrar por hoje. 😊	Vamos dar por terminado o dia de hoje. 😊	"Let's call it a day" is more literal and less metaphorical than the "Cat's out of the bag." Not surprisingly all machine translation output was better. See full review.
I'm head over heels for him.	Estou louca por ele. 😊	Estou louca por ele. 😊	Estou de cabeça erguida para ele. 😞	Eu estou apaixonada por ele 😊	Eu estou completamente apaixonada por ele 😊	Microsoft this time around was the worst, and by far with a terrible, unintelligible translation. See full review.
I get the hang of it.	Eu pego o jeito. 😊	Eu entendo o jeito. 😞	Eu pego o jeito. 😊	Eu entendi 😊	Eu finalmente consegui entender. 😊	All three engines had similar results and consistent ratings from Chat GPT. See full review.
Hit me up.	Bata-me. 😞	Me bata. 😞	Bata-me para cima. 😞	"Me avise" ou "Me ligue" 😊	"Fique à vontade para me avisar" ou "Não hesite em me ligar" 😊	Amazon, Google and Microsoft were huge blunders in this one, completely missing the point of "hit me up" and translating it literally as "beat me up". See full Review.
It really took the wind out of my sails.	Realmente tirou o vento das minhas velas. 😞	Isso realmente tirou o vento das minhas velas. 😞	Isso realmente tirou o vento das minhas velas. 😞	"Realmente me desanimou" ou "Realmente me deixou sem fôlego" 😊	"Realmente me desmotivou ou "Realmente me tirou o ânimo"" 😊	All three engines had terrible, literal translations that make perfect gramatical sense...See full review.
They are two peas in a pod.	São duas ervilhas em uma vagem. 😞	São duas ervilhas em uma vagem. 😞	São duas ervilhas em uma vagem. 😞	Eles são dois pombinhos 😞	Eles são quase idênticos 😊	All three engines had terrible, literal translations that make perfect gramatical sense...See full review.
It takes two to tango.	São precisos dois para dançar o tango. 😞	São precisos dois para dançar o tango. 😞	Quando um não quer dois não brigam. 😊	"É preciso de dois para dançar o tango" ou "É preciso de duas pessoas para dançar o tango" 😞	"Para que haja tango, é preciso de duas pessoas" ou "Sem duas pessoas, o tango não pode acontecer" 😞	Google, Amazon and GPT all failed while Microsoft performed stellarly, perfectly adapting the expression.See full review
It's a piece of cake.	É moleza. 😊	É moleza. 😊	É mamão com açúcar. 😊	É moleza 😊	É muito fácil 😊	Every engine crushed this one. Naturally, b because it was a "Piece of cake" :) See full Review
That costs an arm and a leg.	Isso custa um braço e uma perna. 😞	Isso custa um braço e uma perna. 😞	Isso custa um braço e uma perna. 😞	Isso custa uma fortuna 😊	Isso custa uma fortuna e mais um pouco 😊	Microsoft, Amazon and Google all tanked this one, literally translating the words but losing the meaning in the process ... See full review

Evaluation of GPTs Evaluation

As far as translation quality goes, GPT did a great job with contextualization. In 9 out of 10 sentences, the content was well adapted, intelligible, and conveyed the appropriate meaning. Contrary to the three Machine Translation engines, GPT had no embarrassing “I want to cry” mistakes.

The initial hypothesis was that there would be a big difference in quality between GPT’s first and second iterations, but translation quality was similar in both.

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	70%	70%	50%
Quantitative Analysis Accuracy	60%	60%	40%

Analysis and Key Findings

- Microsoft made bolder choices when it comes to the linguistic adaptation of the idioms.
- GPT had a harder time evaluating Microsoft's metaphorical choices as they departed more.
- Google and Amazon had extremely similar results, only slightly deviating from each other, mirroring each

other's mistakes and metaphorical choices. Microsoft clearly stood out from the two.

- GPT-3 had an easier time with the Qualitative Analysis producing cogent textual analysis (even though with only 70% accuracy).
- Although intelligible GPT's analysis failed to identify in 30% of the cases. This coincided with metaphorical choices that were literal and understandable but deviated from quotidian discourse.
- GPT-3 had a harder time translating the qualitative analysis into a score.

Although broadly speaking scores were 60% accurate, it was difficult to differentiate between similar scores such as a 3 vs. a 4.

- Extreme score divergence from 1 to 5 was easier to understand and more compatible with overall comments suggesting that:
 - Perhaps scoring criteria was not sufficiently calibrated with GPT-3
 - Perhaps binary scoring could be more relevant than gradient scoring
- In one anomalous case Cchat GPT-3 evaluated two similar translations

in radically different ways giving one a 1 and the other a 5 when both of them should have been 1.

- Even though quantitatively Microsoft performed similarly to Google and Amazon, when you get into the nitty gritty of language, Microsoft made bolder choices and provided better results from a qualitative perspective but still was far behind GPT-3 when it came to accuracy and cultural adaptation.

Practical applications and limitations

In this analysis GPT-3 provided superior contextualization and adaptation than previous machine translation models in Brazilian Portuguese. While it's not convenient to replace traditional Machine Translation models with larger models such as GPT-3 due to high computational costs and diminishingdiminhshing marginal gains when it comes to non-metaphorical discourse, GPT-3 can be a powerful human ally when it

comes to providing suggestions and identifying potential mistakes as well as opportunities for improvement.

Linguistic edge cases are amazing because they illustrate so clearly how much is left in so little when it comes to language models. Even though they are “nearly” there, this “nearly” becomes progressively harder to tackle and if not harder, definitely more expensive from a computational perspective.

While none of the engines are reliable enough to replace humans (at least in the conext of this study), GPT-3 shows clear capability of aiding human translators and reviewers in the process of translating and evaluating language.

Spanish

The following analysis was written by our English/Spanish linguistic Nicolas Davila.



Overall all engines struggled with the metaphorical nature of language, often erring in excessive literality.



Output from a Qualitative Perspective

The table below is the synthetic view analysis of our linguistics on 10 English idioms sentences translated into Spanish.

	Google	Amazon	Microsoft	GPT
Success Rate	20%	30%	10%	70%
Awkward Rate	20%	20%	10%	20%
I want to cry	60%	50%	80%	10%

As far as translation quality refers, GPT did an acceptable job and better than the others with contextualization. In 7 out of 10 sentences, the content was intelligible, well formed, and conveyed the appropriate meaning, with low awkward and I want to cry rates.

Although the initial hypothesis was that there would be a big difference in quality between GPT's first and subsequent iterations, translation quality is similar in all of them, 2nd and 3rd iterations sometimes add unnecessary stuff, rising a bit the awkward rate.

The table below contains the **raw data analysis**.

English Sentence Input	Google Output	Amazon Output	Microsoft Output	Chat GPT Output (without any guidance)	2nd Chat GPT iteration "More figuratively"	Overall Human Comments
The cat's out of the bag.	El gato está fuera de la bolsa. 😞	El gato está fuera de la bolsa. 😞	El gato está fuera de la bolsa. 😞	El gato se ha salido del saco. 😞	El secreto ya no es un secreto. 😊	All three machine translations were the same and too literal translations not conveying meaning accurately. GPT's original translation is also a literal one, but when...See full review.
Let's call it a day.	Digamos que ha sido todo por hoy. 😞	Demos por terminado el día. 😞	Terminar la jornada. 😞	Terminemos por hoy. 😞	Pongamos un punto final a esto. 😊	Although this sentence is more literal and less metaphorical than previous one, I have to note that it can be used not only to state that "we should stop working or doing ...See full review.
I'm head over heels for him.	Estoy loca por él. 😞	Estoy perdidamente enamorada de él. 😊	Estoy locamente por él. 😞	Estoy locamente enamorada de él. 😊	Me siento como si estuviera cabeza abajo por él. 😞	Here, the best of all three MT outputs is Amazon's one, as it accurately conveys the meaning; Microsoft's one was the worst, with a terrible translation lacking of the proper....See full review.
I get the hang of it.	Me doy cuenta. 😞	Lo entiendo. 😞	Le entiendo. 😞	Lo tengo entendido. 😞	Creo que lo tengo asido. 😞	""To get the hang of something"" is an idiom for ""to learn how to do something, esp. when it is not simple or obvious."" In this sense, none of MT outputs convey accurately...See full review.
Hit me up.	Golpeame. 😞	Golpéame. 😞	Golpéame. 😞	Mandame un mensaje. 😊	Contáctame cuando tengas tiempo. 😊	All three MT outputs are terribly literal and totally wrong. Although GTP scoring and assessment are consistent and it seems to understand the idiom's meaning,...See full Review.
It really took the wind out of my sails.	Realmente me quitó el viento de las velas. 😞	Realmente me quitó el viento de las velas. 😞	Realmente me quitó el viento de las velas. 😞	Realmente me desanimó. 😊	Realmente me dejó sin aliento. 😞	All MT outputs are terrible and too literal. What's odd is that even though GPT knows the right context when it refers to translation, it's evaluations were all off, rating..See full review.
They are two peas in a pod.	Son dos guisantes en una vaina. 😞	Son dos guisantes en una vaina. 😞	Son dos guisantes en una vaina. 😞	Son dos gotas de agua. 😊	Son como dos uvas del mismo racimo. 😞	In this case all three MT outputs are terrible and literal, and also GPT's scoring and assessments are consistent and accurate. But only the 1st iteration of GPT is accurate and natural...See full review
It takes two to tango.	Se necesitan dos para bailar un tango. 😊	Se necesitan dos para bailar un tango. 😊	Hacen falta dos. 😞	Se necesitan dos para bailar el tango. 😊	Se necesita a dos para que esto funcione. 😊	Considering that there is an equivalent idiom for this in Spanish "Para bailar tango, se necesitan dos"; Both Google's and Amazon's outputs and 1st GPT's iteration are accurate... See full Review
It's a piece of cake.	Es un trozo de tarta. 😞	Es pan comido. 😊	Es un juego de niños. 😊	Es pan comido. 😊	Es muy fácil. 😊	Google's output is the worst one, being it a word by word literal translation. Both Amazon's and Microsoft's outputs, together with all GPT's iterations are good ones...See full Review
That costs an arm and a leg.	Eso cuesta un brazo y una pierna. 😞	Eso cuesta un brazo y una pierna. 😞	Eso cuesta un brazo y una pierna. 😞	Eso cuesta un ojo de la cara. 😊	Eso cuesta una fortuna. 😊	All three MT outputs are too literal and not used frequently in Spanish: Although GPT seems to understand the idioms meaning it gives the same good score and assessment.... See full review

Evaluation of GPTs Evaluation

	Google	Amazon	Microsoft
GPT Qualitative Analysis Accuracy	50%	50%	30%
Quantitative Analysis Accuracy	50%	50%	30%

Analysis and Key Findings

- MT translations were too literal, being Amazon and Google very similar in general terms, and Microsoft being the worst one.
- Google and Amazon had extremely similar results, only slightly deviating from each other, mirroring each other's mistakes and metaphorical choices. Microsoft performed poorly, sometimes producing sentences ill formed and missing some parts of the grammatical construction.
- GPT-3 had an easier time with the Qualitative Analysis producing coherent textual analysis, even though with only 50% accuracy.
- Frequently, GPT-3 qualitative analysis was too general and more restricted to the main and literal meaning of the sentence, without considering subtle details of construction and change in meaning. It seems GPT is not able to catch such differences and to translate them into quantitative scores.
- Also, GPT frequently assigned the same qualitative analysis and high quantitative score to sentences that were grammatically poor constructed, which seems to be a limitation of GPT's model.

GPT-3 had a harder time translating the qualitative analysis into a score. Being it only 50% accurate for Google and Amazon and only 30% accurate for Microsoft. It seems GPT was only measuring if the sentence conveyed the meaning, but no differences in construction or well formation.

Score divergence from 1 to 5 was easier to understand and more compatible with overall comments suggesting that: Perhaps scoring criteria was not sufficiently calibrated for GPT-3 Perhaps binary scoring could be more relevant than gradient scoring Perhaps scoring criteria or GPT model were not considering grammatical issues, but only conveying of meaning

Practical applications and limitations

In this analysis GPT-3 provided superior contextualization and adaptation than previous machine translation models in Latam Spanish. GPT-3 could be a powerful tool in helping humans to improve translations when it refers to provide useful suggestions and opportunities for improvement. But, as far as I can see, it still has certain limitations.

Although larger models such as GPT-3 could be helpful, it is not convenient to replace traditional Machine Translation Models with them, due to higher computational costs due, as when it comes to non-metaphorical text it could diminish marginal gains.

While GPT-3 shows clear capability of aiding human translators and reviewers in the process of translating and evaluating language, cost considerations should be included when evaluating its use for non-metaphorical texts.

Conclusions

- ✧ While ChatGPT's performance can vary depending on the prompt and the specific language, our study suggests that ChatGPT has the potential to produce higher quality translations than traditional MT engines, especially when it comes to handling idiomatic expressions and nuanced language use. However, it is important to note that ChatGPT is far from not making mistakes and still has immense room for improvement, especially when it comes to more complex prompts or language domains.
- ✧ As a translator, ChatGPT was more successful than all tested Machine Translation engines. While languages showed different results, Korean was clearly the outlier with MT quality and GPT quality significantly lower than other languages. In all languages except Korean, ChatGPT had at least a 70% success rate and at most a 90% success rate, performing better than traditional MT. And even in Korean, while scores were low, they still were better than MT engine output.
- ✧ Contrary to Machine Translation, with an LLM, iterations of the same content can improve the output quality. This is key when thinking about integrations because whereas with traditional MT your output will always be the same to your input (unless the engine gets further data or training), with an LLM one can explore several interactions via API in order to optimize feed quality.
- ✧ One advantage of ChatGPT over traditional MT engines is its ability to learn and improve over time, even without additional training data. This is due to the nature of LLMs, which are designed to continually refine their language models based on new input. As such, ChatGPT can potentially offer more adaptive and dynamic translation capabilities, which could be especially useful in scenarios where the language or content is constantly evolving or changing.

✧ Another advantage of ChatGPT is its low cringe rate, which is a significant improvement over traditional MT engines that often produce awkward or inappropriate translations. This could make ChatGPT more acceptable and user-friendly for non-expert users who may not have the same level of linguistic or cultural knowledge as professional translators. However, it is important to note that ChatGPT is not a substitute for human translators, and there is a myriad of cases where the expertise and judgment of a human translator are needed. But ChatGPT's significantly lower cringe rates opens the door for a wider adoption of non-human driven translations.

✧ As an evaluator, ChatGPT's performance was more mixed, with accuracy rates ranging from 30% to 70%. While this suggests that ChatGPT may not be as effective at evaluating other translation engines as it is at suggesting translations, it is possible that

this is due to the complexity and quality of the evaluation prompts, which may require more specialized or contextual knowledge than ChatGPT currently possesses. Further research is needed to explore ChatGPT's potential as an evaluator, as well as its limitations and challenges.

✧ Overall, our study suggests that ChatGPT has promising translation capabilities that are worth exploring further. While it may not be able to replace or bypass human translators entirely, it could potentially offer significant benefits as an aid or pre-translation tool, especially in scenarios where time, resources, or expertise are limited. As with any emerging technology, there are still many challenges and opportunities for improvement, and further research and experimentation will be needed to fully unlock its potential.



THANK YOU!

Schedule a conversation and find out how our translation services and platform will engage your audience globally.

Talk to us