

Part 4: from method to usable system

Human-Centered NLP

Philippe Laban (Salesforce Research)



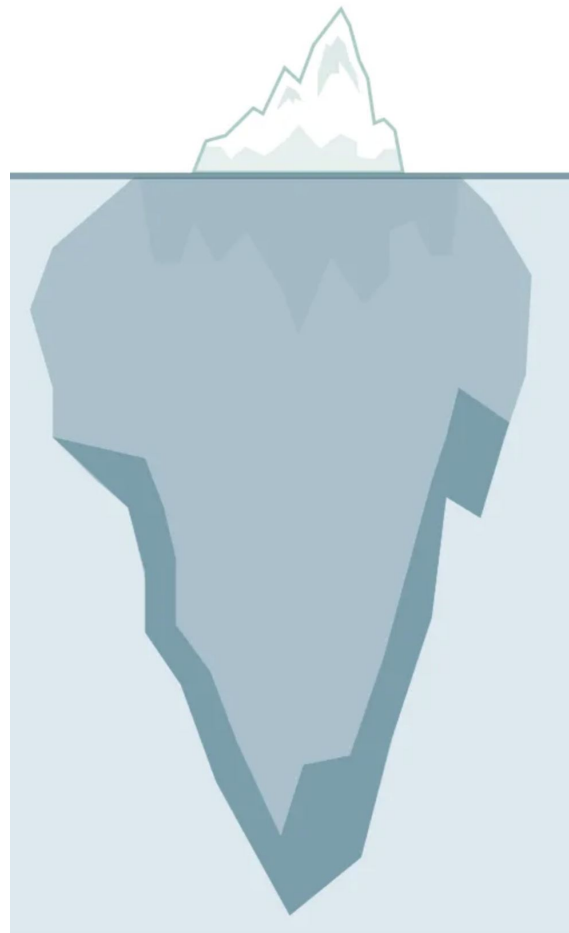
It seems like LLMs can already perform & evaluate text editing tasks...

Is text editing "solved"?

What keeps us from using it in practical settings?

(you might be too shy to ask...)

Going from a method
to a usable system:
Human-Centered NLP



An Intro to Human-Centered NLP

PART 1

Beyond the ML Method

Situation: I have built a cool NLP model and I want to people to use it!

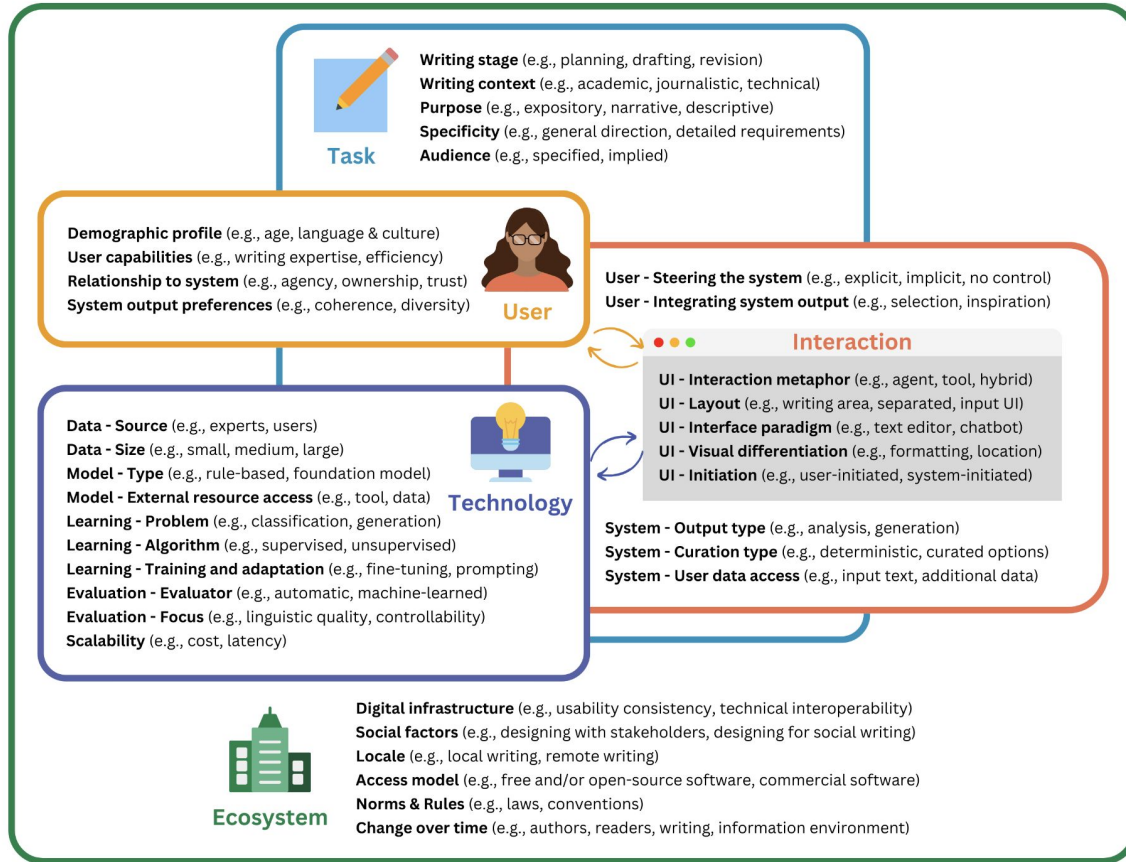
What should I consider when building the system that uses my NLP method?

PART 2

Usability Testing Crash Course

Situation: I have built a system that uses my NLP model. How do if “it works” for users?

A deep-dive into two usability studies of an LLM-based text-editing system.



Technology is only one aspect to consider when building a System.

Important to also consider what the **Task** is, who the **User** is, what the **Interaction** is, and how the **Ecosystem** works.

See this great paper surveying 115 papers in the space!

Task: What is the user trying to accomplish?

Example: consider the Writing Stage of the user.

What stage of writing does the system support?

(1) Idea Generation, (2) Planning, (3) Drafting, or (4) Revision?

Task: What is the user trying to accomplish?

Example: consider the Writing Stage of the user.

What stage of writing does the system support?

(1) Idea Generation, (2) Planning, (3) Drafting, or (4) Revision?



A chat interface might be most adequate for these...

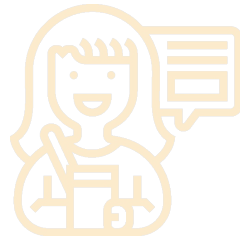
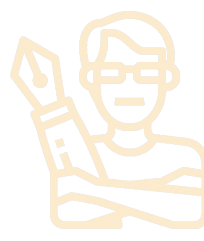
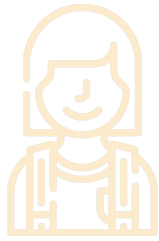
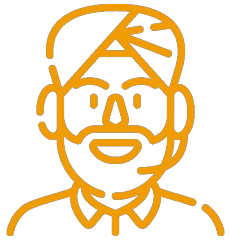
... auto-complete for this

... & a Grammarly-style text editor for this

(see the paper for other task considerations: Writing Context, Purpose, Specificity, and Audience)

User: Who is the target user of the system?

Example: What are the target User Capabilities?

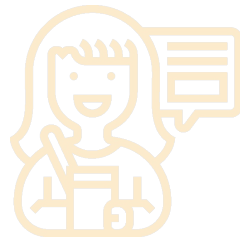
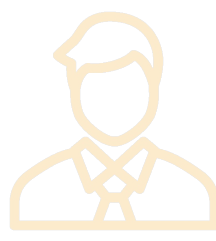
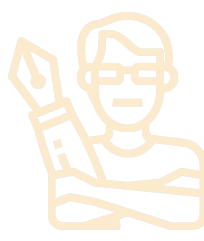
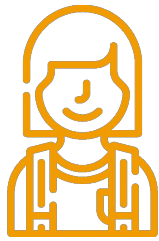
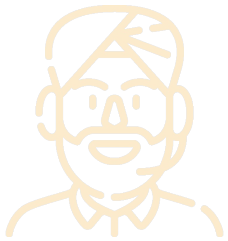


An international
Wikipedia Editor might need help with
American English rules.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

User: Who is the target user of the system?

Example: What are the target User Capabilities?



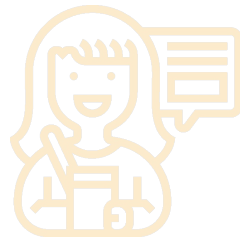
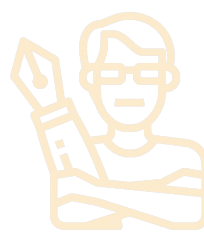
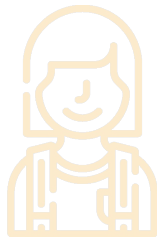
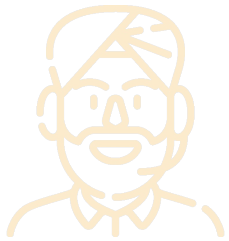
A 6th Grade
Student

might need help spotting
grammar & spelling issues.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

User: Who is the target user of the system?

Example: What are the target User Capabilities?



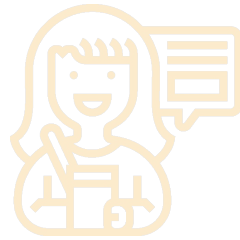
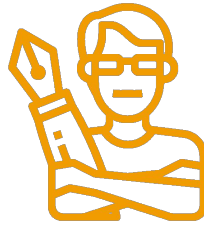
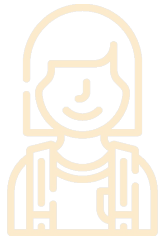
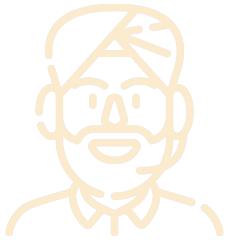
A Ph.D.
Student

might need help making sure
technical terminology is accurate.

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

User: Who is the target user of the system?

Example: What are the target User Capabilities?



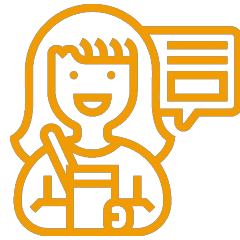
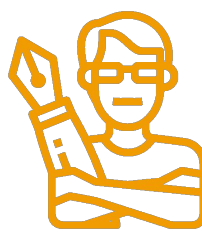
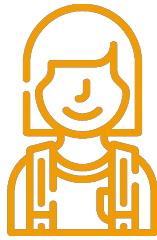
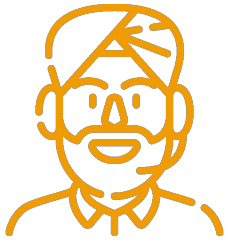
A professional
writer

might not want the system to
make changes affecting their style

(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

User: Who is the target user of the system?

Example: What are the target User Capabilities?



The user's capabilities and needs should be considered during system design!

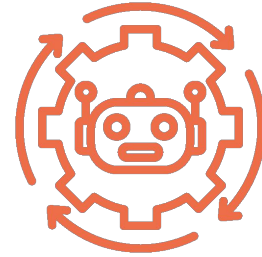
(see the paper for other user considerations: Demographics, Relationship to System, System Output Preferences)

Interaction: How do User, User Interface, and System interact?

Example: How is the system output triggered? (Initiation)



User-Initiated
Reactive



System-Initiated
Proactive

(see the paper for other interaction considerations: Interface Paradigm, Visual Differentiation, Steering, Integration)

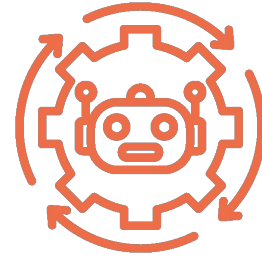
Interaction: How do User, User Interface, and System interact?

Example: How is the system output triggered? (Initiation)



User-Initiated
Reactive

More natural for
Idea Generation & Planning
(ask for what you want)



System-Initiated
Proactive

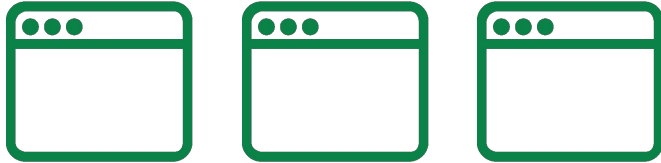
More natural for
Revision, Grammar, Typos
(continuous scan & fix)

(see the paper for other interaction considerations: Interface Paradigm, Visual Differentiation, Steering, Integration)



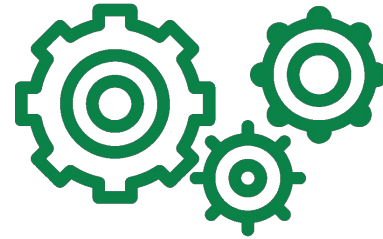
Ecosystem: Does the system fit in the overall ecosystem?

Example: What compatibility issues does the system consider?



Usability
Consistency

Does the system align with
other systems in terms of
usability?



Technical
Operability

With external services (APIs,
tangible world, etc.)

(see the paper for other interaction considerations: Access Model, Social Factors, Locale, Norms, Change over Time)

Task, User, Interaction, Ecosystem

...

What if I just want to build cool NLP models?



Good news: You do not need to “check all the boxes”.

But building useful systems likely requires considering some of these.

Let's see it in action:

InkSync

Executable and Verifiable
Text-Editing with LLMs

Philippe Laban, Jesse Vig, Marti Hearst,
Caiming Xiong, Chien-Sheng Wu
UIST 2024

Motivation: Imagine that you are drafting an email

In the past... you have probably sent an email draft to ChatGPT, asking for feedback on:

- How to change to the (+ formal)
- Fix typos/grammar
- Insert additional details
- Etc.



You

Help me make my email sound more formal and professional. Also add details about things to do in Rome.
...

Subject: Plan your next vacation.

Hey Bob,

How have you been? Are you tired and wanna go on a vacation? I have special invitation to Rome, here are the details.

Experience vibrant cultur and explore the stunning landscapes the place has to offer.

Dont miss out on this fantastc opportunity! Reach out to us now to book your escape. Your dream vacation awaits!

Cheers,
InkSync Agency
...



Motivation: LLM Typical Reply to Rewriting

The LLM typically replies with a new version of the document.

Pros:

- Ready-to-use
- Convenient to copy-paste

Cons:

- Editing suggestions are implicit
- Not easy to accept subset of edit suggestions

We surveyed knowledge workers: most who use LLMs for rewriting frequently accept LLM rewrites verbatim. Most look at suggestions, and decide which to include in their editor.



ChatGPT

Subject: Plan Your Next Vacation in Rome

Dear Bob,

I trust this message finds you in good health and high spirits. If you have been contemplating a getaway, we have an exclusive invitation for you to explore the captivating city of Rome. Here are the details.

Rome is a city steeped in history, offering a plethora of attractions and experiences. While you are here, you can visit iconic landmarks such as the Colosseum, the Vatican City, and the Roman Forum. Explore the charming streets of Trastevere, savor exquisite Italian cuisine in quaint trattorias, and indulge in the world-famous gelato.

Do not miss out on this fantastic opportunity to embark on an unforgettable journey. Please feel free to reach out to us at your earliest convenience to discuss the details and book your dream vacation in Rome.

We look forward to assisting you in creating a truly memorable experience.

Best regards,

[Your Name]

InkSync Agency



Motivation: Reverse-Engineering the LLM Output

Using a string-alignment algorithm (Levenshtein), we can recover the exact edits proposed by the LLM: the characters **inserted** and **deleted**.

This makes the editing suggestions more **explicit**.

But such string alignment visualization is not common in chat interfaces, and taking editing actions based on this would require manual work, that could be error-prone.

Subject: Plan ~~y~~our ~~n~~ext ~~v~~acation **in Rome**.

~~Hey~~**Dear** Bob,

~~How have you been?~~ **I trust this message finds you in good health and high spirits.** ~~Are you tired and wanna go on a vacation?~~ **If you have been contemplating a getaway, † have special** **we have an exclusive** invitation to **explore the captivating city of Rome.**
~~h~~**Here** are the details.

~~Experience vibrant cultur and explore the stunning landscapes the place has to offer.~~ Rome is a city steeped in history, offering a plethora of attractions and experiences. **While you are here, you can visit iconic landmarks such as the Colosseum, the Vatican City, and the Roman Forum. Explore the charming streets of Trastevere, savor exquisite Italian cuisine in quaint trattorias, and indulge in the world-famous gelato.**

~~Don't~~ **not** miss out on this fantastic opportunity **to embark on an unforgettable journey!** Please feel free to **R**each out to us ~~now~~ **at your earliest convenience** to discuss the details and book your ~~escape~~ **dream** vacation in Rome.

~~Your dream vacation awaits!~~ **We look forward to assisting you in creating a truly memorable experience.**

~~Cheers~~**Kind Regards,**
[Your Name]

InkSync Agency

Motivation: What if we ask for explicit edits directly?

We could ask the LLM to provide its feedback in bullet-point form? Adding this to the prompt:

Do not rewrite the document. Instead, list out exactly all the editing changes you would make to the document.

Sorry for the small font, answer is **much longer**.

Pros:

- LLM provides explanations/reasons.
- Transparent: know how many edits are suggested.

Con: Still hard to integrate, requires manual work, so nobody does this.

 ChatGPT

Here are the editing changes I would make to your email to make it sound more formal and professional:

1. Change the subject line to be more descriptive and professional:
Subject: Discover Rome: Your Next Vacation Destination
2. Address the recipient more formally:
Change "Hey Bob," to "Dear Bob,"
3. Improve the opening sentence to sound more professional:
Change "How have you been?" to "I hope this message finds you well."
4. Make the invitation more formal:
Change "I have special invitation to Rome, here are the details." to "We are pleased to extend an exclusive invitation for you to visit Rome. Please find the details below."
5. Enhance the description of Rome:
Change "Experience vibrant cultur and explore the stunning landscapes the place has to offer." to "Immerse yourself in the vibrant culture and explore the breathtaking landscapes that Rome has to offer."
6. Correct the spelling and grammar errors:
 - Change "cultur" to "culture."
 - Change "fantastc" to "fantastic."
7. Include more information about Rome's attractions:
Add a brief paragraph about popular things to do in Rome, such as visiting the Colosseum, exploring the Vatican City, trying authentic Italian cuisine, and strolling through historic neighborhoods like Trastevere.
8. Encourage action in a professional manner:
Change "Dont miss out on this fantastic opportunity! Reach out to us now to book your escape." to "Don't miss out on this fantastic opportunity! We encourage you to contact us at your earliest convenience to secure your dream vacation."
9. Sign off more formally:
Change "Cheers, InkSync Agency" to "Best regards, InkSync Agency"

By making these changes, your email will appear more formal and professional while also providing additional information about things to do in Rome to entice the recipient.





In InkSync, we go a step further, and ask the LLM to generate executable edits, in a standardized edit language.

Executable edits are first-class objects in the text editor, which the user can review, accept & dismiss in one click.

Executable Edit Language - Example Edit

```
{  
  "original_text": "Hey Paul",  
  "replace_text": "Dear Paul",  
  "replace_all": "o",  
  "component": "chat",  
  "new_info": "o"  
}
```

Title: Come see Singapore!

Hey Dear Paul,

's and you well. I am

✓ Accept Suggestion

✗ Dismiss Suggestion

InkSync Demo!

InkSync has **four components** that can produce executable edits:



CHAT



COMMENT



MARKERS



BRAINSTORM

<https://inksync.salesforceresearch.ai/>

InkSync Demo!

InkSync has **four components** that can produce executable edits:



CHAT



COMMENT



MARKERS



BRAINSTORM

In this talk: we'll look at two of them, see paper/demo for other two.

InkSync Demo



CHAT

1. The user asks for editing help
 2. The LLM replies:
 - 2a. In plain text (in the chat)
 - 2b. With a list of executable edits (viewed in the editor)
- Because Chat is LLM-based, it can handle a broad range of editing tasks.*

The screenshot displays the InkSync application interface. On the left is a dark purple sidebar with the 'InkSync' logo at the top. Below the logo are several buttons: '+ New', 'My Email' (which is highlighted in blue), 'Tweet Thread', 'Untitled Doc', and 'Audit' (with a power icon). The main content area is titled 'My Email' and shows an email draft with the following text: 'Subject: Followup on call – Vacation to Rome?', 'Hey Alice, It was great talking with you last week, I'm excited to help you plan your trip to Rome, Italy.', 'There you'll experience vibrant cultur and explore the stunning landscapes the place has to offer.', 'You'll also get to eat delicious dishes, as Rome is quite famous for food.', 'Dont miss out on this fantasc opportunity! Reach out to us now to book your escape. Your dream vacation awaits!', and 'Cheers, InkSync Agency'. On the right side of the interface, there are settings and chat controls. At the top right is a 'Settings' gear icon. Below it is a 'View – Inline' toggle switch which is turned on. Further down are three buttons: 'Chat' (with a chat icon), 'Comment' (with a comment icon), and 'Verify' (with a verify icon). A light gray chat bubble contains the text 'Don't hesistate get help in the chat.'. At the bottom right, there is a text input field with the placeholder 'Send message' and a green send button with a white arrow.

InkSync Demo



MARKERS

1. The user creates a Marker once. (e.g.: *look out for Typos*)

2. During editing, Markers continuously suggests executable edits, running in the background.

Markers = proactive
Chat = reactive

The screenshot displays the InkSync application interface. On the left is a dark purple sidebar with the 'InkSync' logo, a '+ New' button, and a list of documents: 'My Email' (selected), 'Tweet Thread', and 'Untitled Doc'. At the bottom of the sidebar is an 'Audit' button. The main content area shows an email draft titled 'My Email' with the subject 'Subject: Followup on call – Vacation to Rome?'. The body text includes 'Hey Alice, It was a pleasure speaking with you last week. I am eager to help you plan your trip to Rome, Italy.' and 'There you'll experience vibrant culture and explore the stunning landscapes and must-see monuments such as the Colosseum, St. Peter's Basilica, and Trevi Fountain.' A red bracket marker is positioned under the word 'Italy'. Below this, the text continues: 'You'll also get to eat delicious dishes, as Rome is known for its rich cuisine such as pasta Carbonara, Roman pizza – a thin, crispy pizza with various toppings – and gelato.' and 'Don't miss out on this fantastic opportunity! Reach out to us now to book your escape. Your awaiting dream vacation!'. On the right side, there is a settings panel with a 'Settings' gear icon, a 'View – Inline' toggle switch (which is turned on), and a 'Markers' section. The 'Markers' section lists two active markers: 'Typos' (with a red pencil icon) and 'Professional' (with a yellow pencil icon). Below the markers is an 'Add Marker' button. At the bottom of the right panel, there is a 'Chat' section with buttons for 'Chat', 'Comment', and 'Verify'. A chat message is visible, partially obscured, mentioning 'Colosseum, St. Peter's Basilica, and Trevi Fountain will make the trip sound even more exciting. Here are my suggestions. Take a look!'. Below the chat message, it says '1 chat suggestion accepted.' At the very bottom right, there is a 'Send a message' input field and a green arrow button.

How does InkSync fit in the Design Space?

Aspect	Dimensions	Tags
--------	------------	------

How does InkSync fit in the Design Space?

Aspect	Dimensions	Tags
Task	Writing Stage	Drafting and Revision (not Idea Generation & Planning)

How does InkSync fit in the Design Space?

Aspect	Dimensions	Tags
Task	Writing Stage	Drafting and Revision (not Idea Generation & Planning)
User	Demographic Profile	Focused on Knowledge Workers (writing daily)
User	Relationship to System	Agency (easy to accept/dismiss individual suggestions), Trust (Factuality Checking)

How does InkSync fit in the Design Space?

Aspect	Dimensions	Tags
Task	Writing Stage	Drafting and Revision (not Idea Generation & Planning)
User	Demographic Profile	Focused on Knowledge Workers (writing daily)
User	Relationship to System	Agency (easy to accept/dismiss individual suggestions), Trust (Factuality Checking)
Interaction	UI — Interface Paradigm	Text Editor with Chat on the Side
Interaction	UI — Visual Differentiation	Formatting (Red/Green Styling)
Interaction	UI — Initiation	Both Reactive (Chat) and Proactive (Markers)

How does InkSync fit in the Design Space?

Aspect	Dimensions	Tags
Task	Writing Stage	Drafting and Revision (not Idea Generation & Planning)
User	Demographic Profile	Focused on Knowledge Workers (writing daily)
User	Relationship to System	Agency (easy to accept/dismiss individual suggestions), Trust (Factuality Checking)
Interaction	UI — Interface Paradigm	Text Editor with Chat on the Side
Interaction	UI — Visual Differentiation	Formatting (Red/Green Styling)
Interaction	UI — Initiation	Both Reactive (Chat) and Proactive (Markers)
Ecosystem	Model Access	Does not allow for model change
Ecosystem	Usability Consistency	Visual Resemblance to other Text Suggestion systems

In summary: All dimensions might not be relevant to every project. But considering where a project fits within the space can help situate the work & come up with ideas.

An Intro to Human-Centered NLP

PART 1

Beyond the ML Method

Situation: I have built a cool NLP model and I want to people to use it!

What should I consider when building the system that uses my NLP method?



PART 2

Usability Testing Crash course

Situation: I have built a system that uses my NLP model. How do if "it works" for users?

A deep-dive into two usability studies of an LLM-based text-editing system.



Advice #1: Consider the Task & Participants Carefully

- ✓ Select participants that would find the task “relevant”.
- ✗ Recruit on MTurk for a task that requires specialized skills.
- ✓ Think carefully of total task duration:
 - Enough time for meaningful interaction & feedback
 - Not too long to avoid burnout / skipping
- ✓ Plan for time to train/onboard participants on how to use the interface.

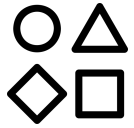
InkSync Study 1: Task Setup

Task: Participants are given an email template from the “*InkSync Travel Agency*”

They have 6 minutes to:

- Objective A: Make sure it sounds professional / doesn't have typos
- Objective B: Customize it to a travel destination / customer persona

They will use different kinds of interfaces with and without assistance.



Advice #2: Select Broad Study Conditions



Make sure to select **baselines** and “**oracle**” settings when possible.



Don't just compare variations of your system.



Map potential differences in conditions to research questions.



Don't overdo it. More conditions = more participants.



Decide on study design:

- Between-subjects: Each participant interacts with **one** condition
- Within-subjects: Each participant interacts with **all** conditions

InkSync Study 1: Study Conditions

Six conditions:

Manual Editing
(Baseline)

Non-Executable Chat
(Baseline)

Chat Only

Markers Only

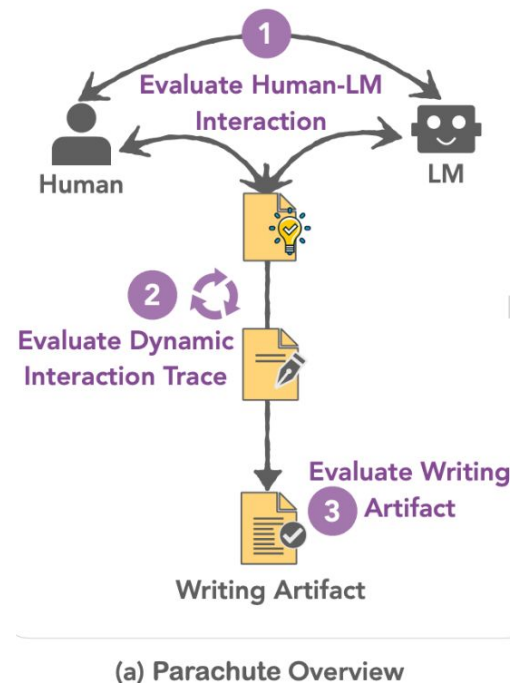
Local Only

InkSync 4-Components

- Recruit ~100 participants.
- Each completing 3 sessions (in random order) -> total 300 editing sessions
- Intro Material (2-min) + 3 x 6-min editing + 2-min survey + 5-min buffer
(you always need to include buffer time, unexpected things happen)

Advice #3: Include Redundancy in Evaluation Signal

- ✓ Evaluate Interaction Logs
- ✓ Evaluate “Product/Artifact” of Study
- ✓ Ask Participants for Preference/Opinion
- ✗ Don't: Treat *Participants as Annotators*.
You will throw away a lot of good signal.



InkSync Study 1: Evaluation Setup

INTERACTION LOGS

Editing Efficiency
(Levenshtein)

Component Usage
Popularity

ARTIFACT ANALYSIS

Objectives A/B
Scoring

Hallucination Analysis

Divergence Analysis

PARTICIPANT OPINION

Likert Usability
Questions

Condition Preference
Ranking

Redundancy = confirmation

InkSync Study 1: Example Completed Task

Initial Email Template - Study 1 (1 of 3)

Subject: Plan your next vacation.

Hey [Customer's Name],

How have you been? Are you tired and wanna go on a vacation? I have special invitation to [City Destination], here are the details.

Experience vibrant cultur and explore the stunning landscapes the place has to offer.

Dont miss out on this fantastc opportunity! Reach out to us now to book your escape. Your dream vacation awaits!

Cheers,
InkSync Agency

Gloria is
traveling with
2 children

Destination:
Egypt

Subject: ~~Your~~Plan The Next ~~Unforgettable~~ Getaway! Trip You Won't Forget to Egypt!

HeyDear [Customer's Name]Gloria,

How's life treating you? Are you Ffeeling exhausted and in need of an escapes? Well, do we have a surprise-an exciting opportunity for you! We're extending a special invitation just for you to experience the trip of a lifetime to Egypt. You can visit the stunning Pyramids of Giza, one of the original seven wonders of [City Destination]: the ancient world! Prepare-to You will surely be amazed!

Get ready Prepare to immerse yourself of-in a world of rich culture and behold the breathtaking landscapes that awaits you. It's a place where dreams come true! memories are created and is an ideal destination for families. There are engaging experiences tailored for all ages, such as camel-riding around the Giza Plateau or a family cruise down the Nile River.

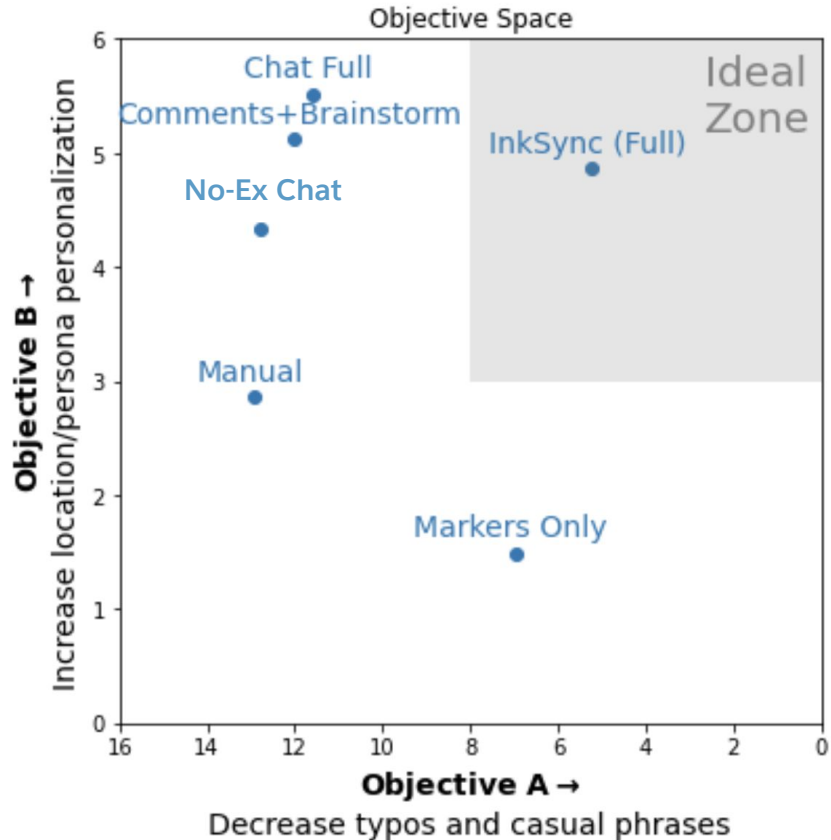
Imagine yourself wandering through the streets, immersing yourself in local tradition, and discovering hidden gems that will make your back home friends envious: visiting local markets, experiencing new culinary delights, snorkeling in the Red Sea, and even hot air ballooning in Luxor.

Don't let Seize this incredible opportunity slip through your fingers without delay. Reach out to us now to secure your spot for an extraordinary vacation in Egypt that you'll cherish forever. Your perfect getaway, complete with a taste of Egypt's delicious cuisine with its delicious dishes like Koshari, Molokhia, and Ta'meya, is just a call away!

CheersKind Regards,
InkSync Agency

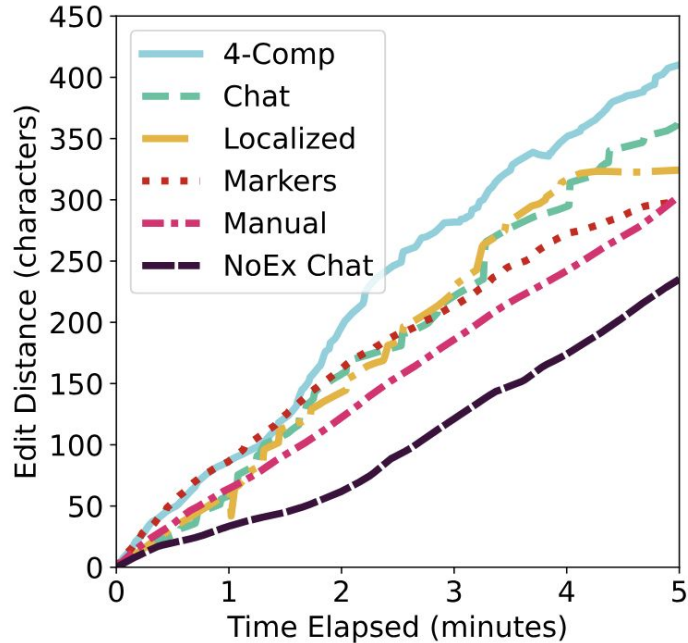
Legend:  Manual Editing  Markers  Chat  Comment  Brainstorm

InkSync Study 1: Objectives A/B Scoring



- “Markers”-like interactions help with Objective A: constantly identifying & remove typos / casual phrases.
- “Chat”-like interactions (even non-executable) help with Objective B: customization / idea generation.
- Combining both (InkSync Full) leads to the best of both: components are complementary.

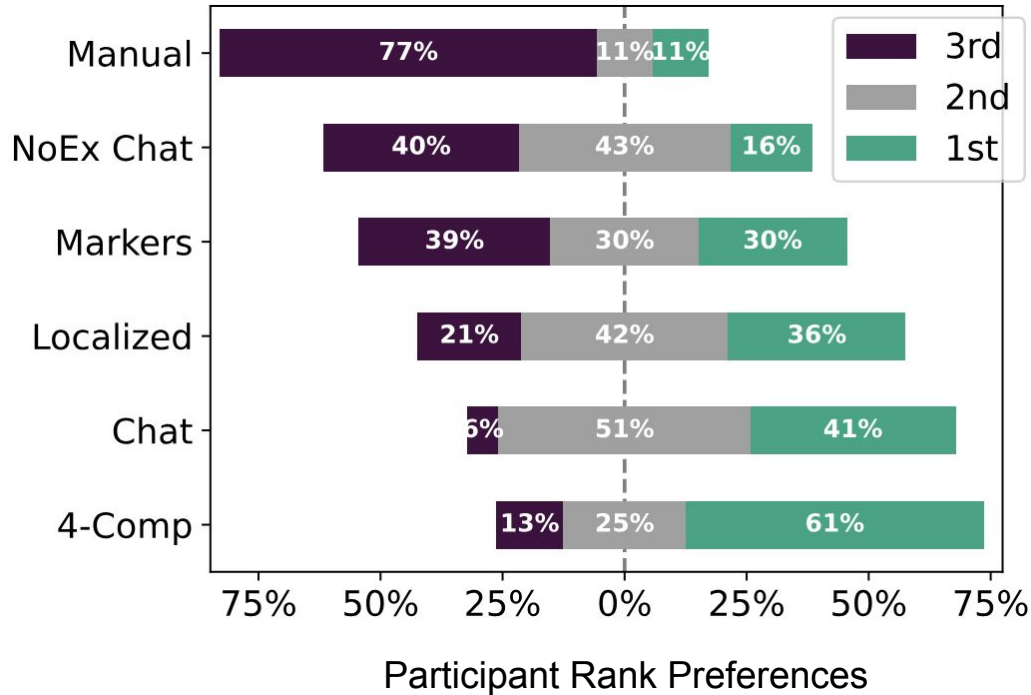
InkSync Study 1: Editing Efficiency



Edit Efficiency Analysis

- Participants with executable edits lead to faster/more efficient editing.
- NoEx Chat even slower than Manual (!) -> divided attention, more effort required.
- 4-Comp (Full interface) leads to fastest editing -> complementary components.

InkSync Study 1: Condition Preference Ranking



Participants ranked the 3 interfaces they used.

4-Comp most preferred, followed by Chat. NoEx Chat & Manual least preferred.

InkSync Study 1: Hallucination Analysis

Manual analysis of ~300 editing sessions reveals:

- 28% of edited emails contain a minor factual error
 - *e.g., saying that Paris is an affordable European city*
- 15% of edited emails contain a major factual error
 - *e.g., recommending a visit to a Zoo that closed in 2017*

To mitigate such risks, we implement a framework that builds on top of executable edits:

WARN - VERIFY - AUDIT



When the LLM generates an executable edit, it must populate a *new_info* key: does the edit introduce new information or not?

When new information is introduced, there's a risk of hallucination and the user is warned.

(a manual analysis confirms GPT4 is ~97.5% accurate at this task)

Executable Edit Language - Example Edit

```
{  
  "original_text": "Hey Paul",  
  "replace_text": "Dear Paul",  
  "replace_all": "o",  
  "component": "chat",  
  "new_info": "o"  
}
```

InkSync Demo



WARN &



VERIFY

When a warning is shown, the user is given an option to **verify** the edit.

To verify the edit, the LLM generates **search engine queries**, that the user can visit to assess the edit's veracity.

Verification is human-in-the-loop by design.

The screenshot displays the InkSync interface. On the left is a dark purple sidebar with the 'InkSync' logo at the top. Below the logo are buttons for '+ New', 'My Email' (highlighted in blue), 'Tweet Thread', and 'Untitled Doc'. At the bottom of the sidebar is an 'Audit' button with a magnifying glass icon. The main content area shows an email draft titled 'My Email'. The email text reads: 'In Rome, you'll experience vibrant culture and explore the stunning landscapes and must-see monuments such as the Colosseum, St. Peter's Basilica, and Trevi Fountain. You'll also get to eat delicious dishes, as Rome is known for its rich cuisine such as pasta Carbonara, Roman pizza -- a thin, crispy pizza with various toppings -- and gelato. To secure your escape, we encourage you to reach out to us at your earliest convenience. Your dream vacation awaits! Kind Regards, InkSync Agency'. A small warning icon is visible in the top left corner of the email content area. On the right side of the interface is a 'Settings' panel with a gear icon. Below it are toggle switches for 'View -- Inline' (which is turned on) and 'Chat' (which is turned off). Under the 'Chat' section, there are buttons for 'Chat', 'Comment', and 'Verify'. A chat window is open, showing a message: 'even more exciting. Here are my suggestions. Take a look!'. Below the message, it says '1 chat suggestion accepted.' and there is a green button labeled 'add emojis'. Another message follows: 'Sure, how about these emoji placements? Don't hesitate to let me know if you want any changes.' Below this, it says '12 chat suggestions accepted.' At the bottom of the chat window is a 'Send a message' input field and a green arrow button.

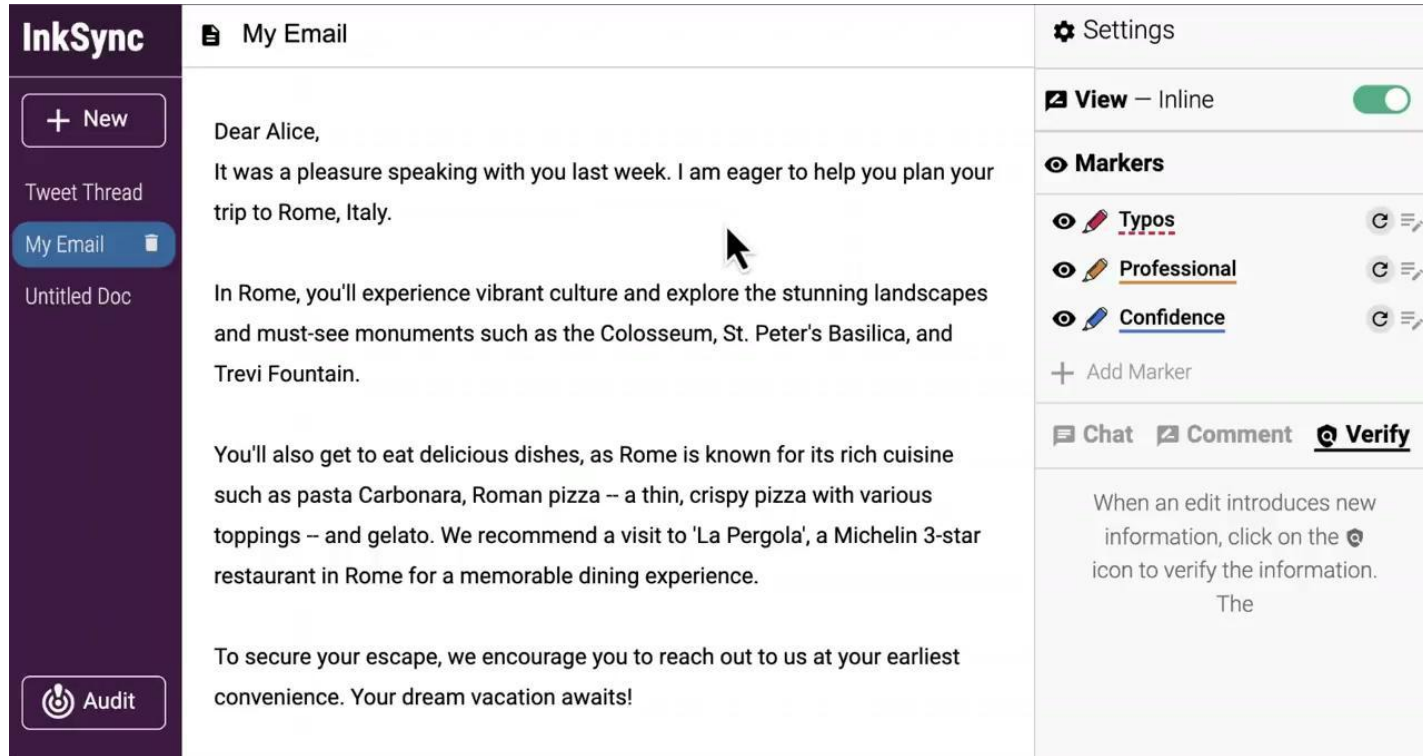
InkSync Demo AUDIT




Because edits are **executed**, we can trace each character's origin (LLM-gen vs. human-written).

We can trace characters during editing sessions, and design a view-only **auditing interface**.

The auditor can verify edits in the interface, performing a final check before the doc is sent/published.



The screenshot displays the InkSync auditing interface. On the left is a dark purple sidebar with the 'InkSync' logo at the top, a '+ New' button, and a list of documents: 'Tweet Thread', 'My Email' (selected), and 'Untitled Doc'. At the bottom of the sidebar is an 'Audit' button with the InkSync logo. The main content area shows an email draft titled 'My Email' with the following text: 'Dear Alice, It was a pleasure speaking with you last week. I am eager to help you plan your trip to Rome, Italy. In Rome, you'll experience vibrant culture and explore the stunning landscapes and must-see monuments such as the Colosseum, St. Peter's Basilica, and Trevi Fountain. You'll also get to eat delicious dishes, as Rome is known for its rich cuisine such as pasta Carbonara, Roman pizza – a thin, crispy pizza with various toppings – and gelato. We recommend a visit to 'La Pergola', a Michelin 3-star restaurant in Rome for a memorable dining experience. To secure your escape, we encourage you to reach out to us at your earliest convenience. Your dream vacation awaits!'. A mouse cursor is positioned over the word 'Fountain'. On the right side, there is a 'Settings' section with a 'View – Inline' toggle switch turned on. Below that is a 'Markers' section containing three markers: 'Typos' (with a red squiggly line icon), 'Professional' (with a yellow pencil icon), and 'Confidence' (with a blue pencil icon). Each marker has a refresh icon and a list icon. At the bottom of the markers section is an 'Add Marker' button. Below the markers is a 'Chat' section with 'Chat', 'Comment', and 'Verify' buttons. A text box below the 'Verify' button contains the instruction: 'When an edit introduces new information, click on the  icon to verify the information. The'.

InkSync Study 2: Study Conditions

Objective: Study whether Warn-Verify-Audit can assist participants in detecting and avoiding hallucinations.

Four conditions:

No Framework
(Baseline)

Warn-Verify

Audit

Warn-Verify-Audit



Advice #4: Adapt the System for the Study



Problem: errors “only” occur in in 15–30% of documents, so not every participant would see a hallucination.



Solution: modify the system to produce **more hallucinations** during the study. Now each participant will encounter factual errors.



A usability study is not a product review.



Control the environment as much as you can.



Advice #5: Prepare to Deal with Cheaters

Unavoidable: some % participants don't do the task well, either because:
(1) they don't care, (2) they don't understand.



Make it easy to spot unengaged participants.



Don't: "Force" participants to complete task (they'll find another way to cheat)



Expect to filter out 10-20% of unengaged participants. Filter all data, don't cherry-pick.



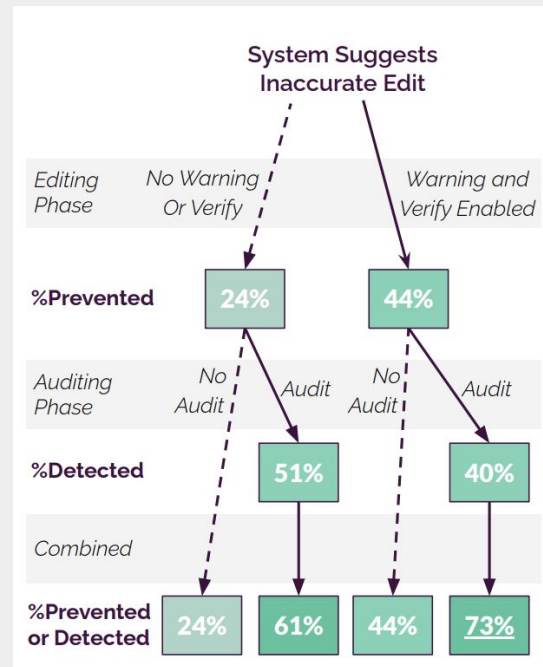
Don't: withhold payment of unengaged participant. It's a part of the game.

InkSync Study 2: Filter out any participant that doesn't run any verification during the study -> indicative of unengaged participant.

InkSync Study 2 High-Level Results

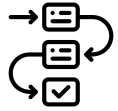
Does Warn-Verify-Audit work?

We ran a second usability study focused on the framework:



Edit- and Audit-time verification are complementary & both help detect and prevent factual errors.

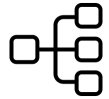
Recap of Usability Study Advice



#1: Consider the Task & Participants Carefully



#2: Select Broad Study Conditions



#3: Include Redundancy in Evaluation Signal



#4: Adapt the System for the Study

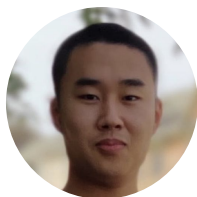


#5: Prepare to Deal with Cheaters

Questions? (don't be shy!)

Some seed questions:

- Do Usability ever scale? What about statistical significance? Don't you always need 1000+ participants??
- My lab can't afford paying for participants, what should I do?
- What is your favorite use-case for text-editing?
- I'm a young researcher, what problem should I tackle?



Yao Dou



Philippe Laban



Claire Gardent



Wei Xu



<https://acl2024-text-generation-tutorial.github.io/>

