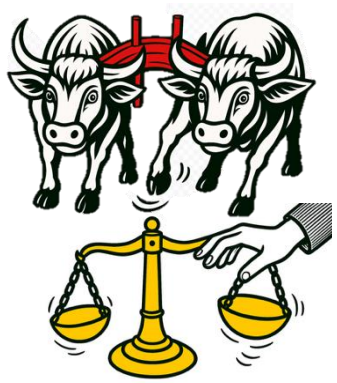
Taking a turn for the better?



Pivoting and pivotal moments in consequential conversations

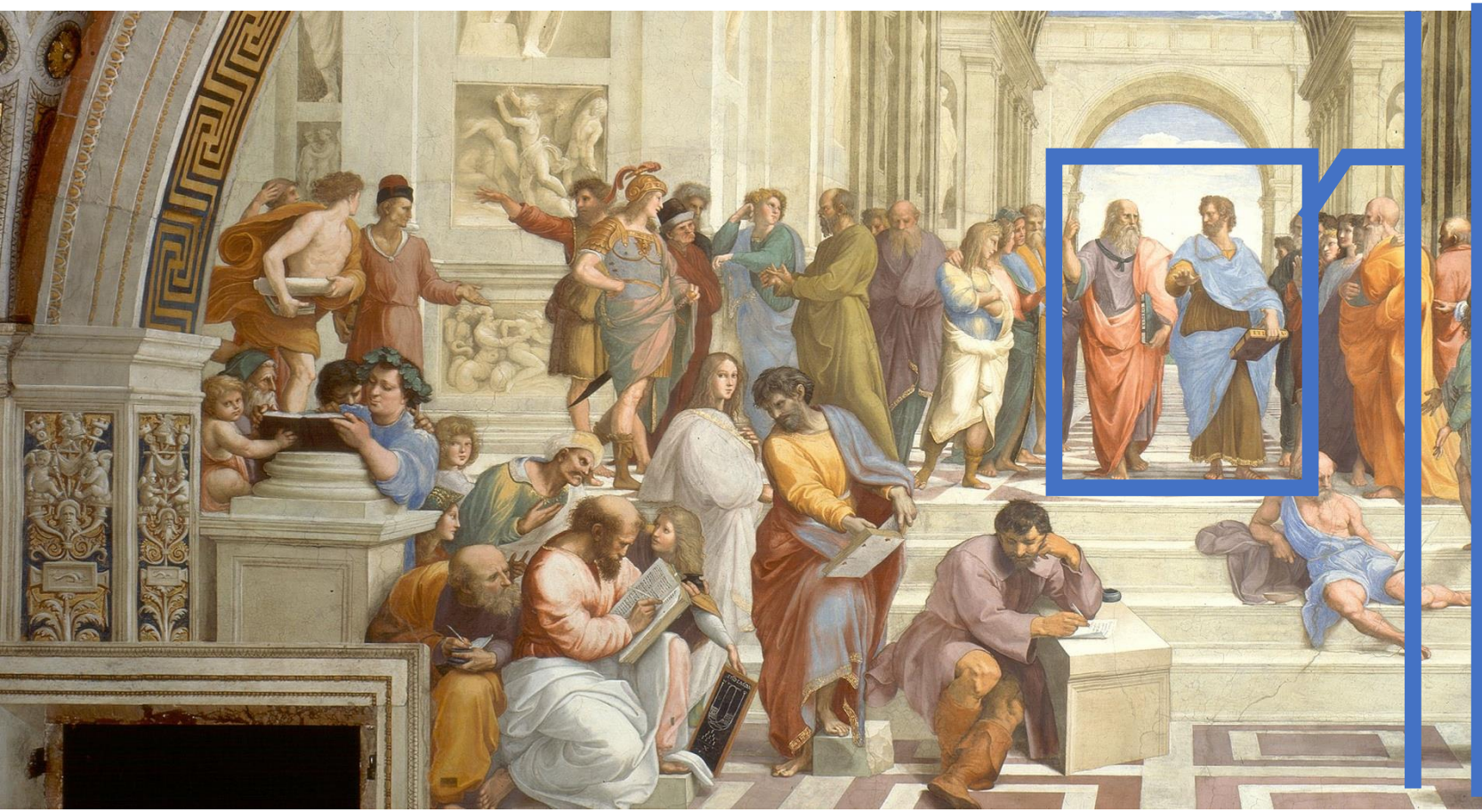
Lillian Lee
Cornell University

Joint work at EMNLP-f. '24, ACL '25, with:
Vivian Nguyen
Sang Min Dave Jung
Thomas D. Hull
Cristian Danescu-Niculescu-Mizil ("D-N-M")

Language-based interactions abound

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound



Face to-face

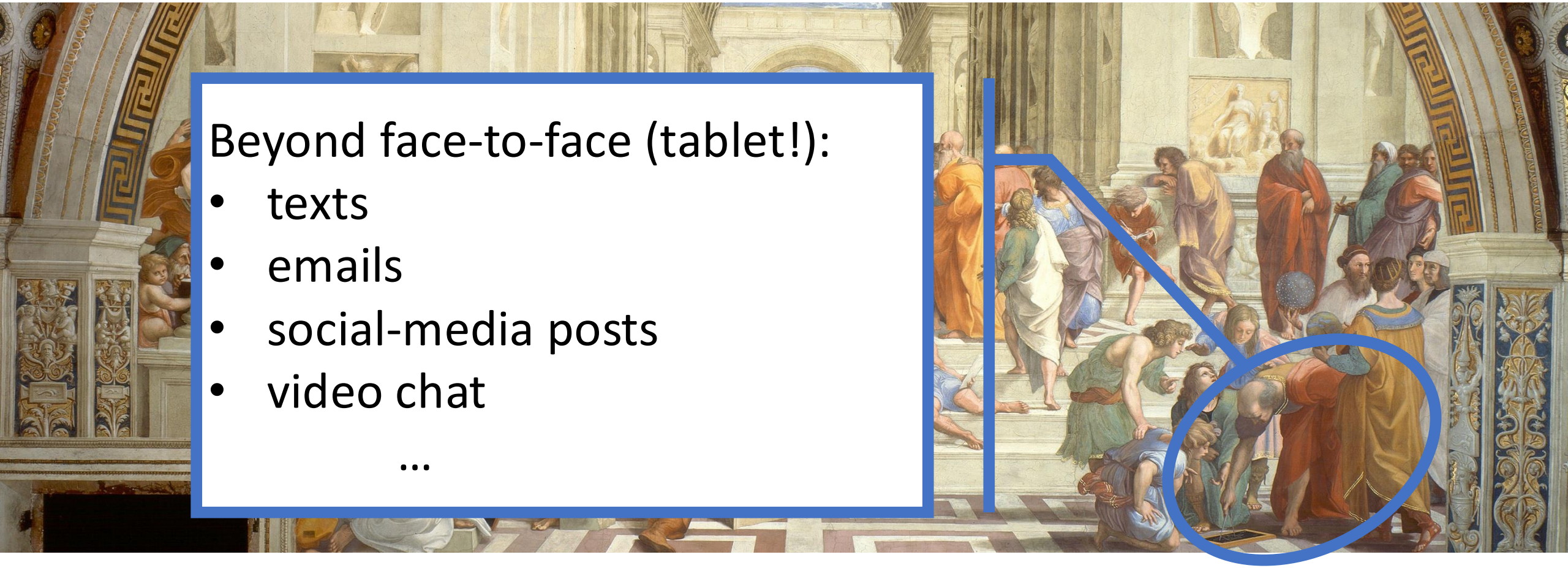
- advising/tutoring
- persuasion
- negotiation
- decision making
- …

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound



Beyond face-to-face (tablet!):

- texts
- emails
- social-media posts
- video chat

…

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound

Message spread & alteration:

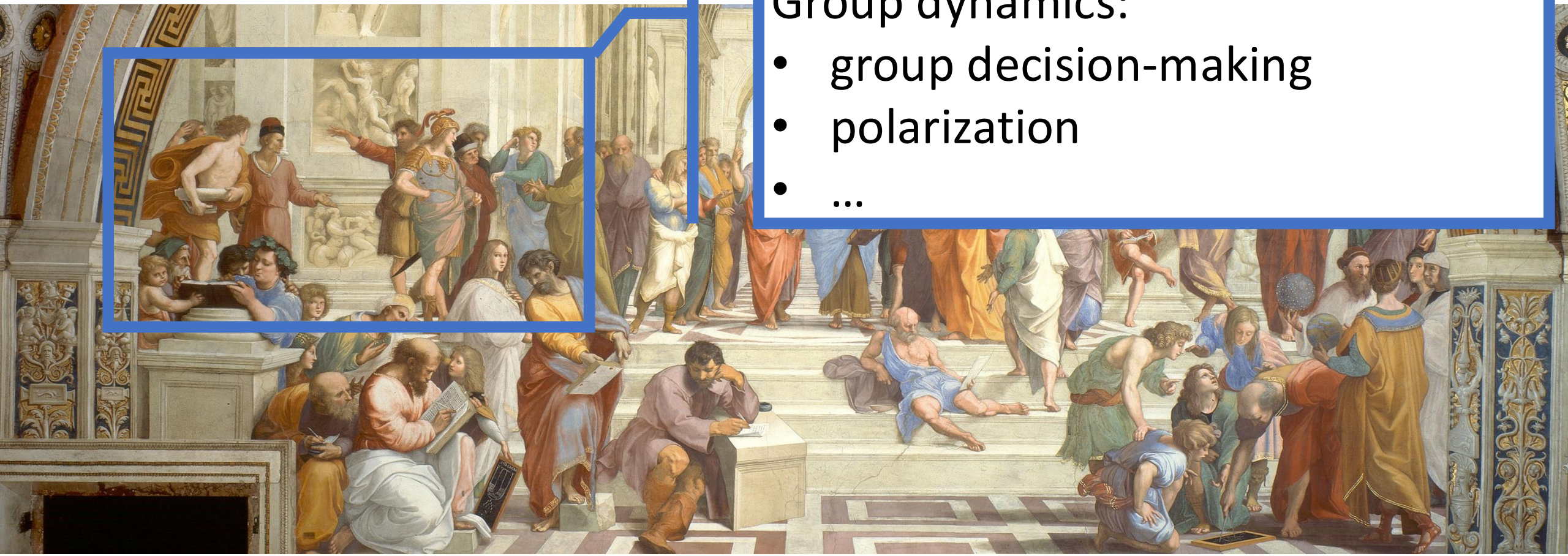
- retweeting
- memes
- rumors and their effects on people's opinions

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound



Group dynamics:

- group decision-making
- polarization
- …

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Language-based interactions abound

Original image: Raphael. https://en.wikipedia.org/wiki/The_School_of_Athens

Conversation outcomes and trajectories

Example: Wikipedians discuss whether to ***keep*** or ***delete*** certain articles.

Outcome:
An official's final decision.

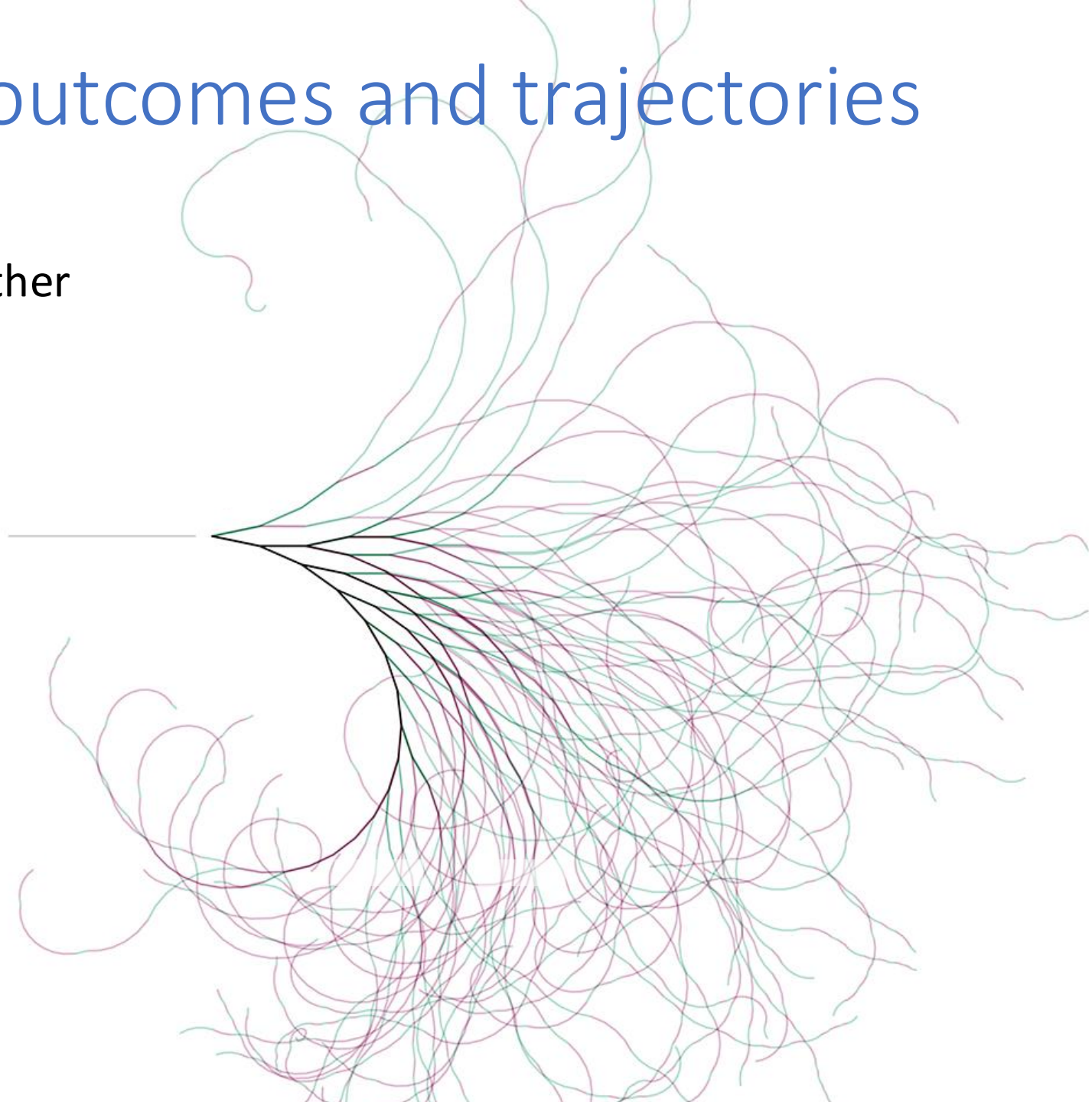
Conversation outcomes and trajectories

Example: Wikipedians discuss whether to ***keep*** or ***delete*** certain articles.

Outcome:
An official's final decision.

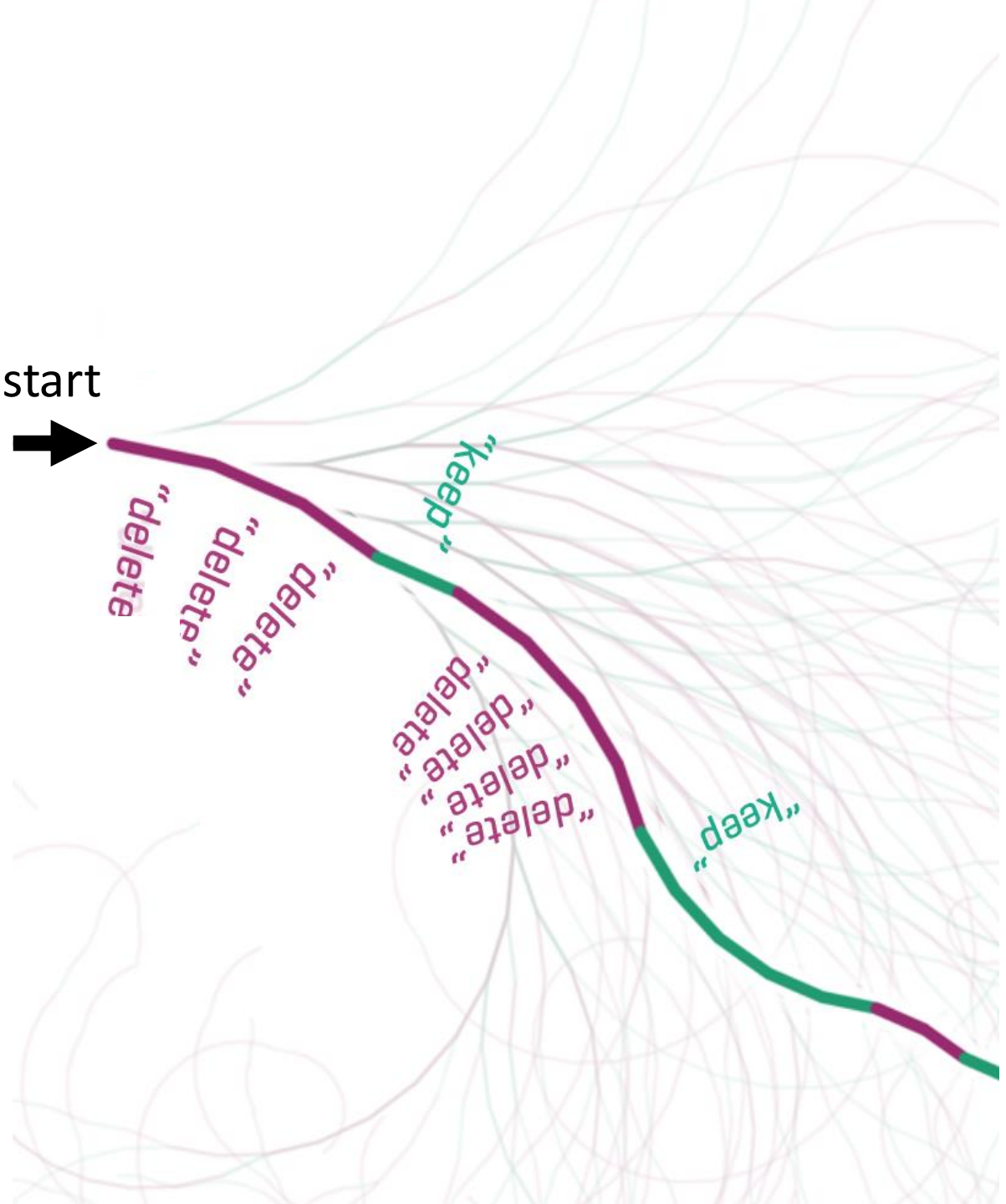
Visualization: trajectories of 100 long discussions.

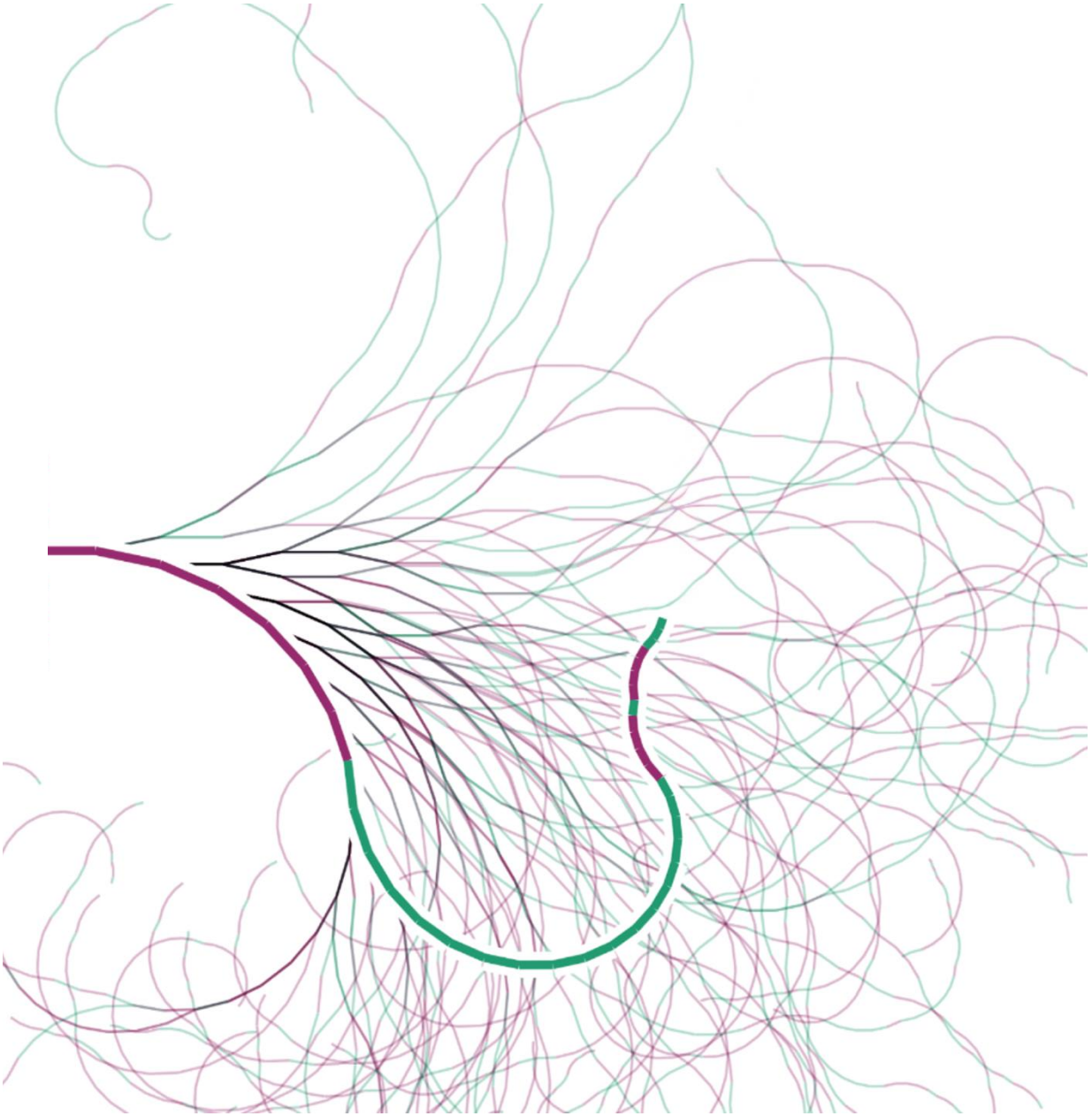
Credit: Stefaner, Taraborelli, & Ciampaglia, https://notabilia.net/



Each "tendril" = sequence of utterances in a particular discussion.

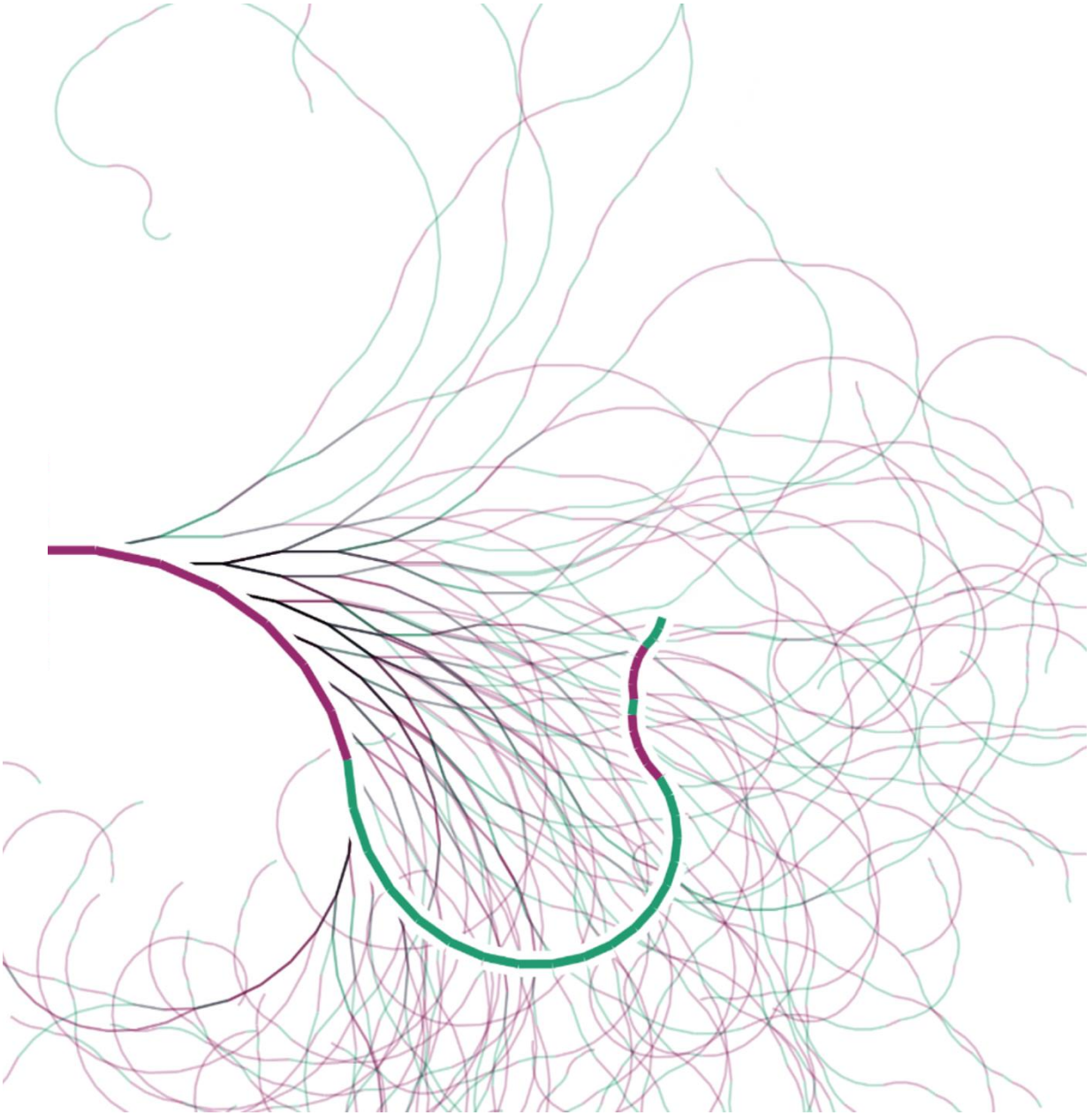
- Next person recommends ***delete***? Tendril turns ***down.***
- An argument for ***keep*** instead? Tendril turns ***up***.





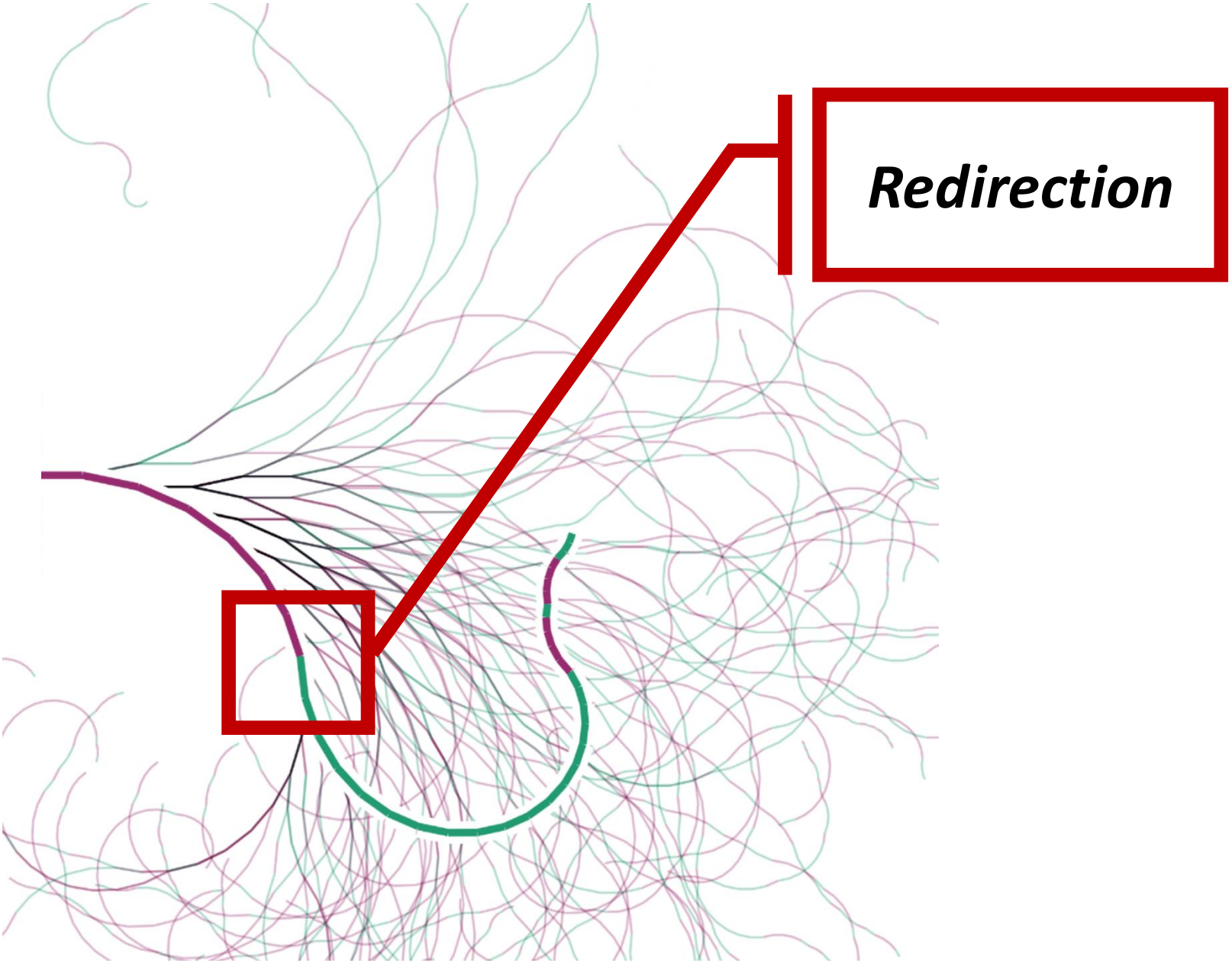
Visualizations by Moritz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia.

Are there "**key moments**" w.r.t final outcome?



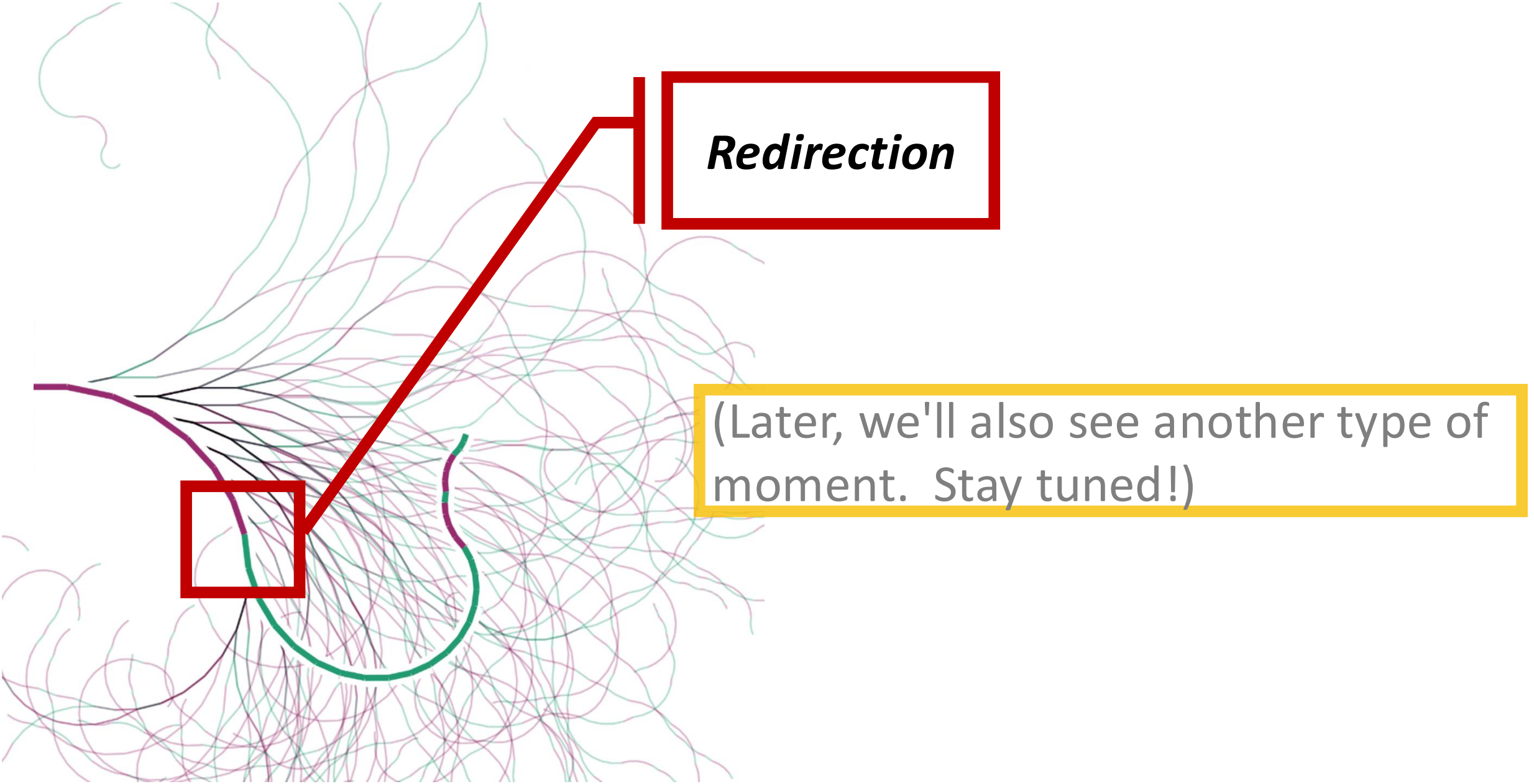
Visualizations by Moritz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia.

Are there "**key moments**" w.r.t final outcome?



Visualizations by Moritz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia.

Are there "**key moments**" w.r.t final outcome?



(Later, we'll also see another type of moment. Stay tuned!)

In this talk: mental-health therapy and crisis counseling

Importance

Immediate use case

Broader frontiers

Slide design adapted from Vivian Nguyen

In this talk: mental-health therapy and crisis counseling

Importance

Immediate use case

Broader frontiers

Globally, >1B people have a mental-health condition [WHO '25].

Huge potential impact of better therapy = conversations.

Outcomes: patient/client improves, or no?

Slide design adapted from Vivian Nguyen

In this talk: mental-health therapy and crisis counseling

Importance

Globally, >1B people have a mental-health condition [WHO '25].

Huge potential impact of better therapy = conversations.

Outcomes: patient/client improves, or no?

Immediate use case

Train therapists/counselors to recognize and react appropriately.

We have active collaboration with provider platforms

Broader frontiers

Slide design adapted from Vivian Nguyen

In this talk: mental-health therapy and crisis counseling

Importance

Globally, >1B people have a mental-health condition [WHO '25].

Huge potential impact of better therapy = conversations.

Outcomes: patient/client improves, or no?

Immediate use case

Train therapists/counselors to recognize and react appropriately.

We have active collaboration with provider platforms

Broader frontiers

Applications to your favorite interaction domain? (Be creative about notion of "outcome".)

- Code: https://convokit.cornell.edu/, the Cornell Conversational Analysis Toolkit

Slide design adapted from Vivian Nguyen

Redirection intuition

"Two decades of empirical research have consistently linked the quality of the **alliance** between therapist and client with **therapy outcome**."

[Horvath '01. emph. added]

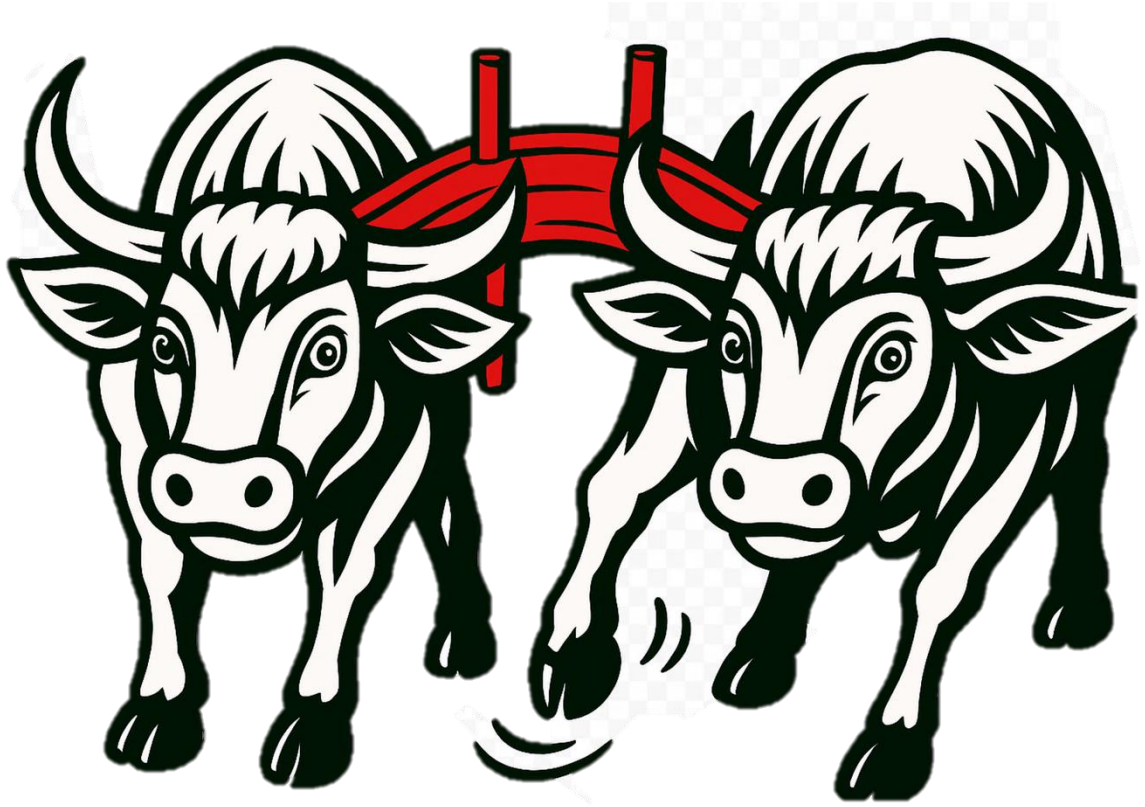
Redirection intuition (cont.)
The psychotherapist and therapy recipient are a **team.**

- They shoulder a burden together.
- They are in it for the long haul.

Changing conversation direction is often necessary to satisfy participant goals.

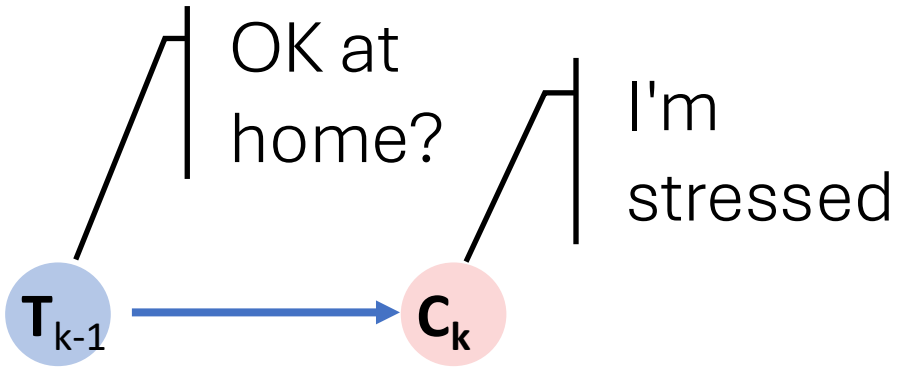
It is a **joint action:**

- If one attempts to redirect …
- … the other must "accept" the attempt for the pivot to occur.



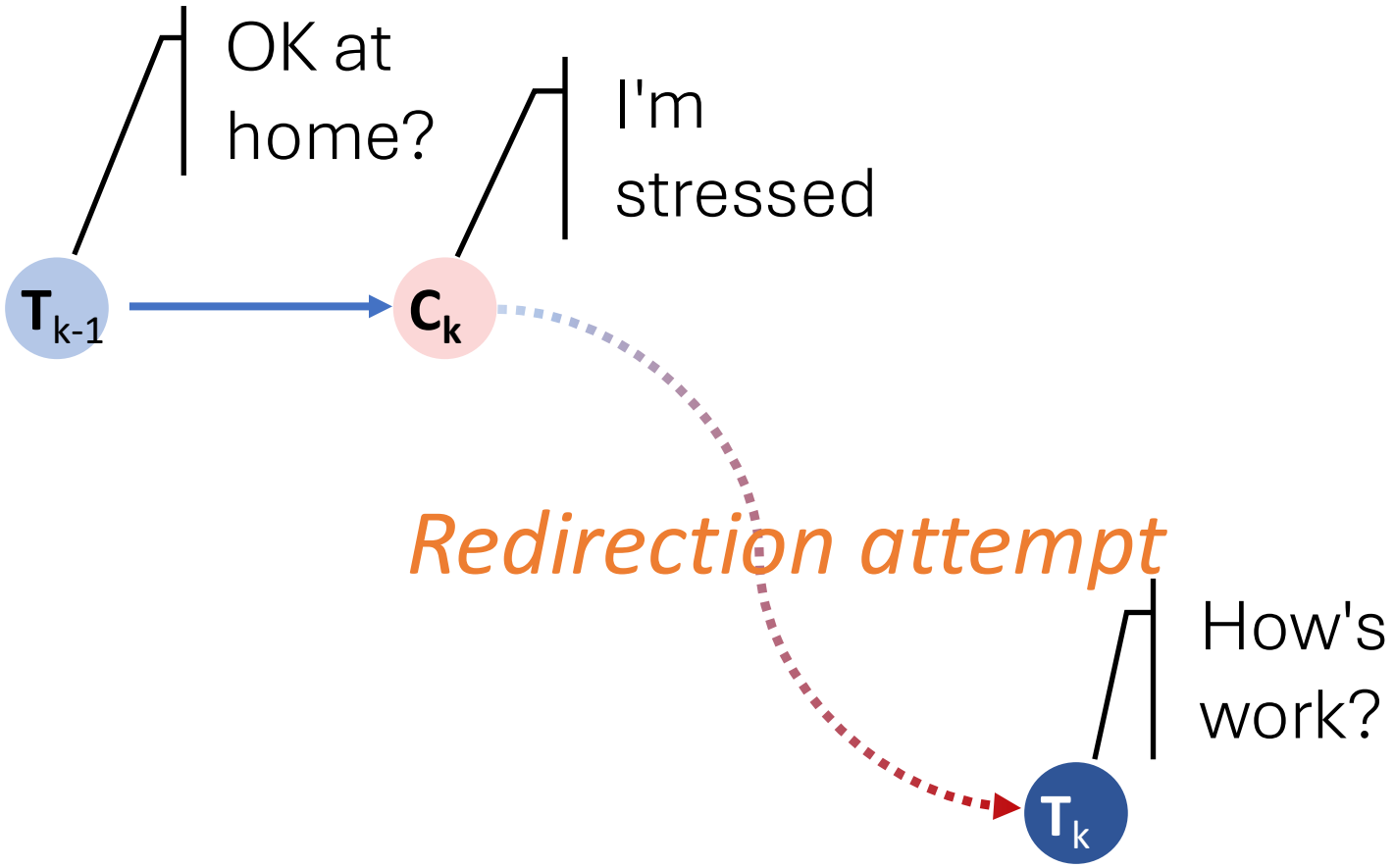
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



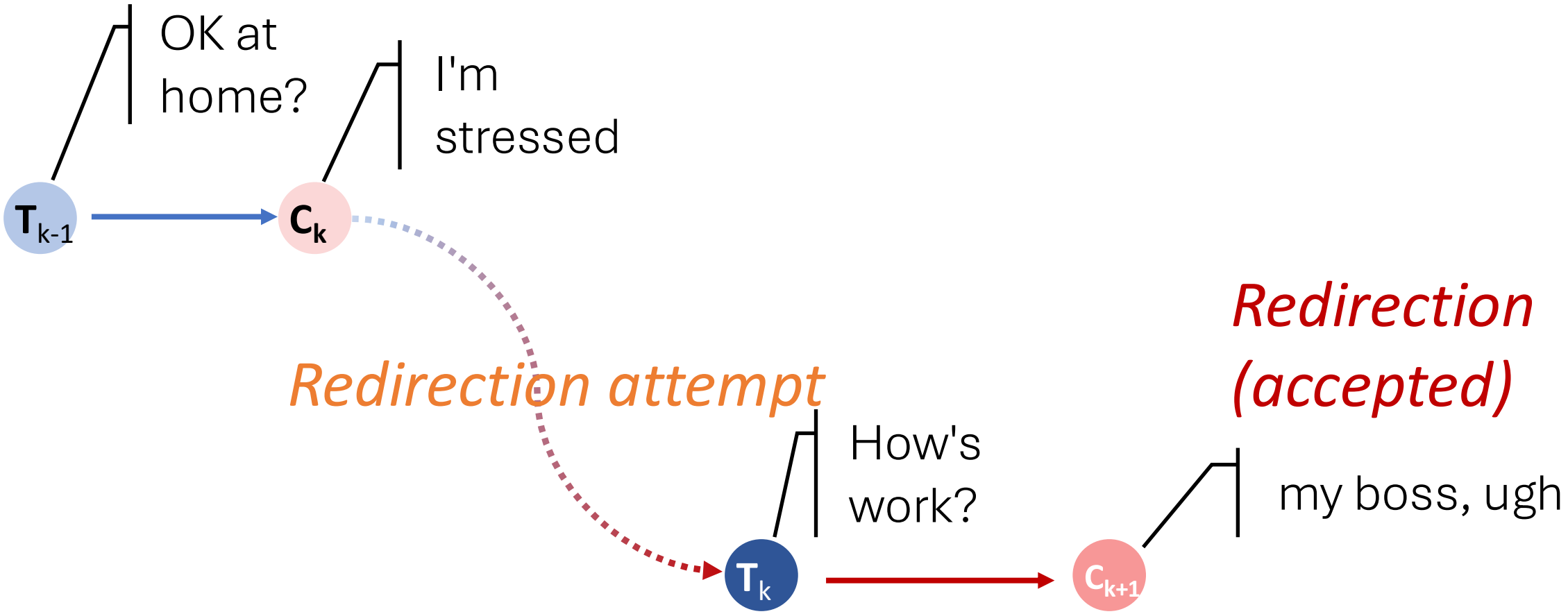
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



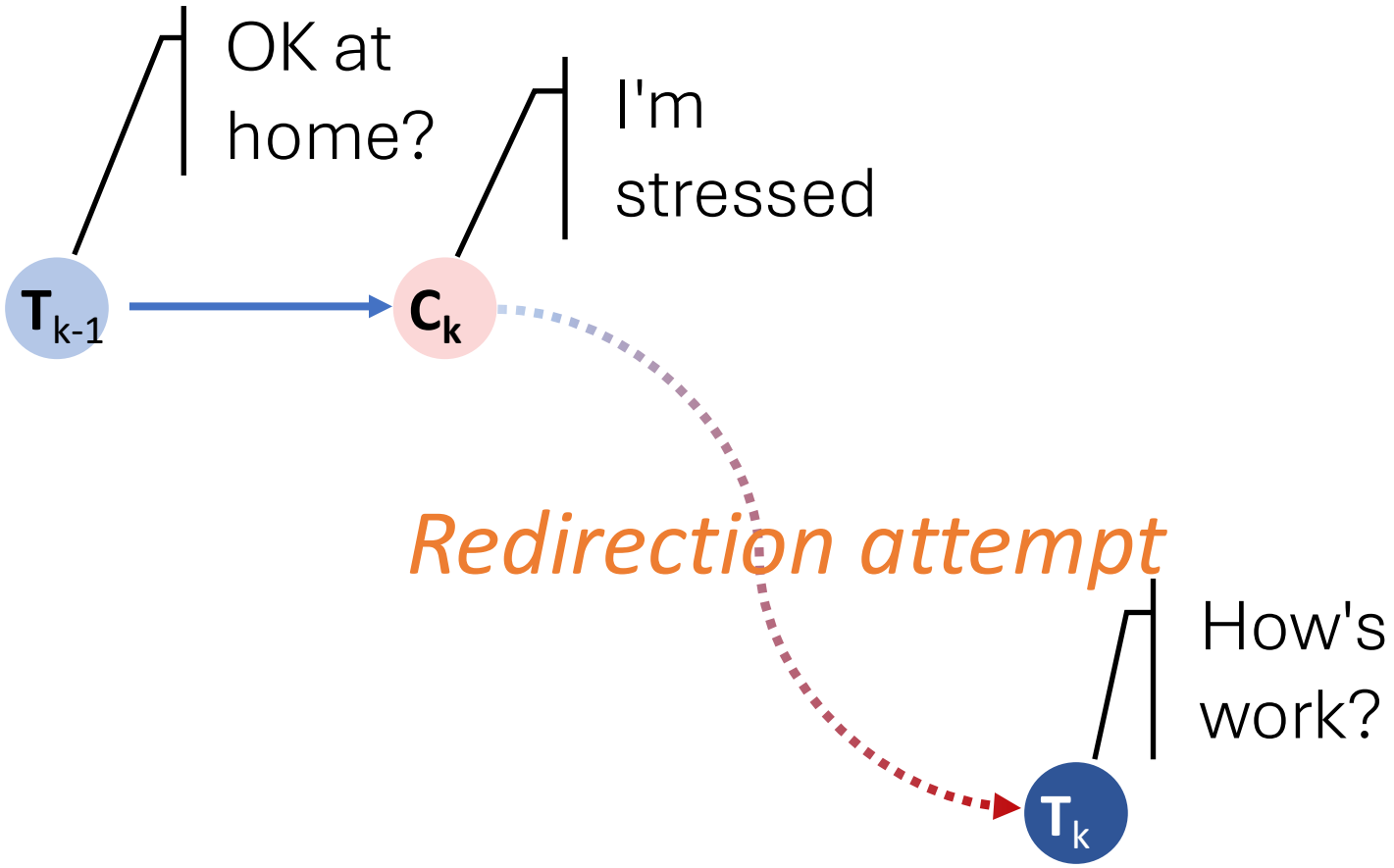
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



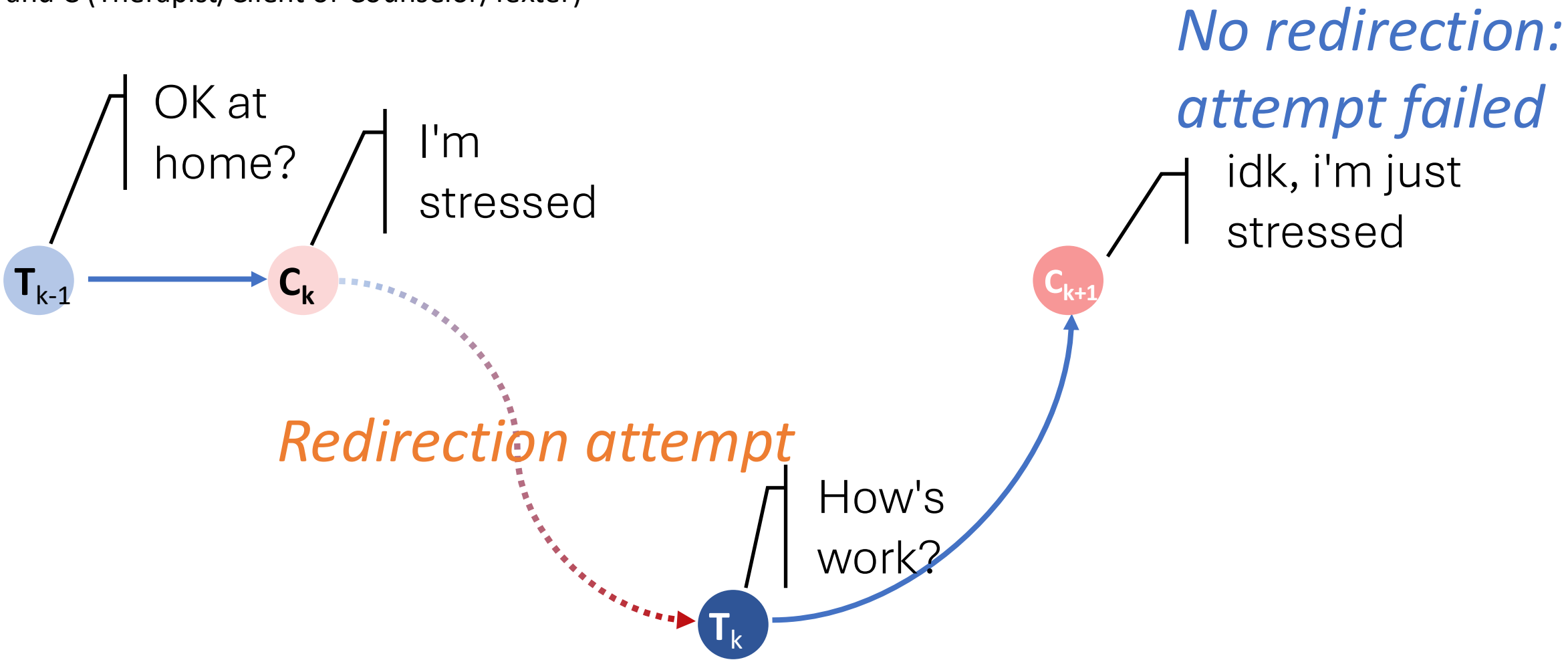
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



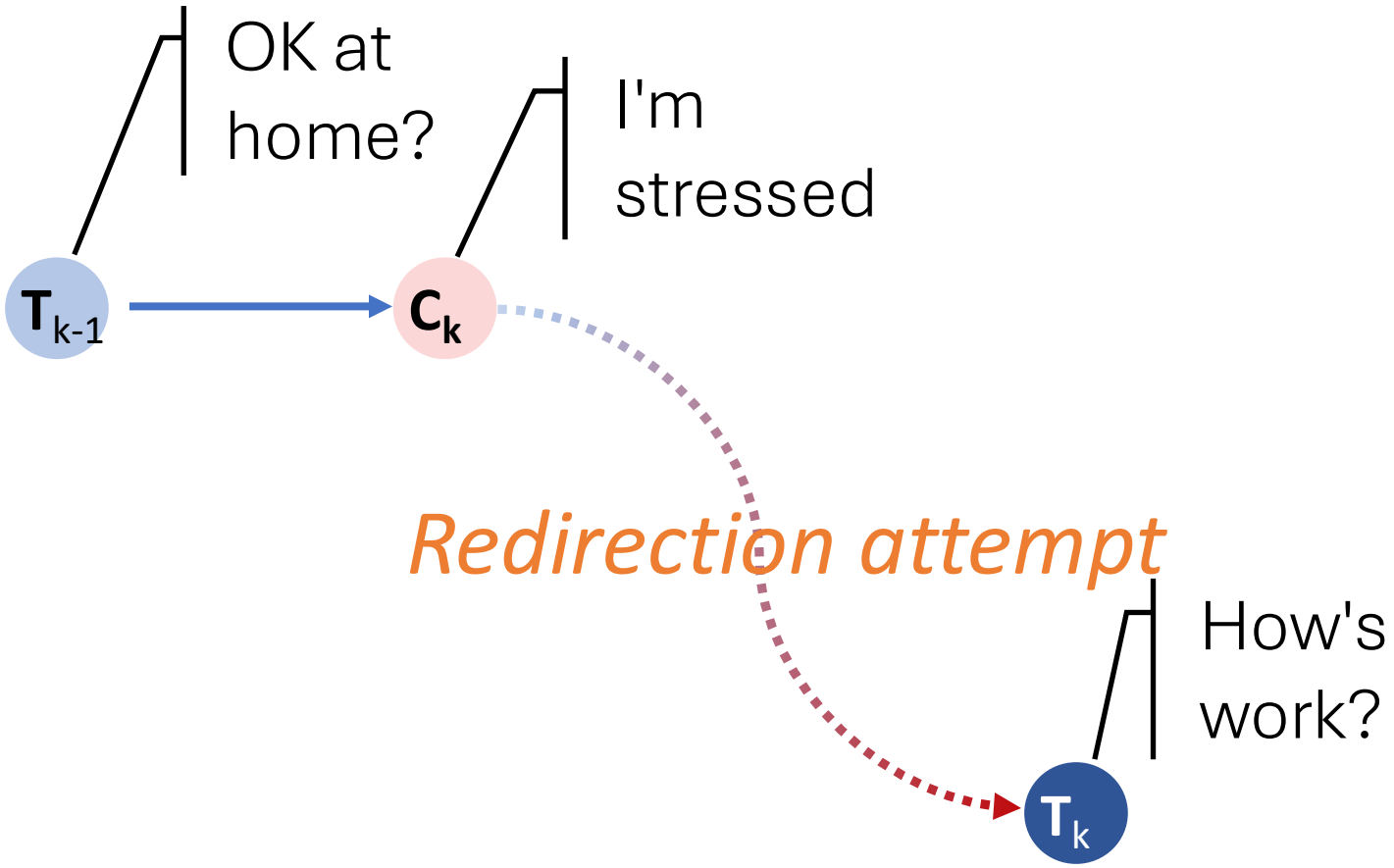
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



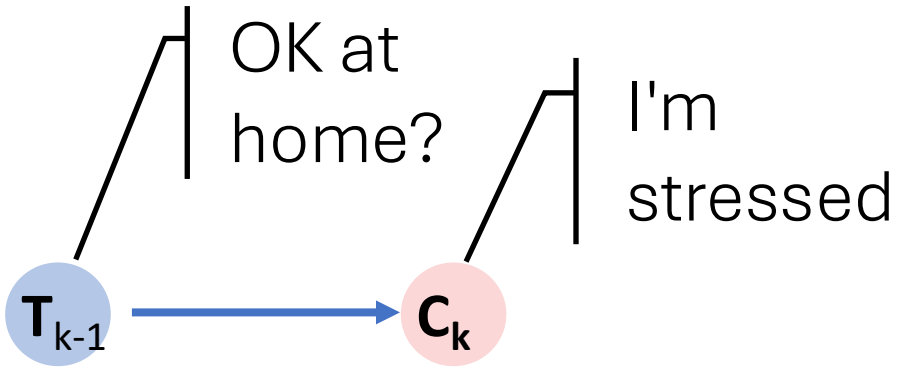
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



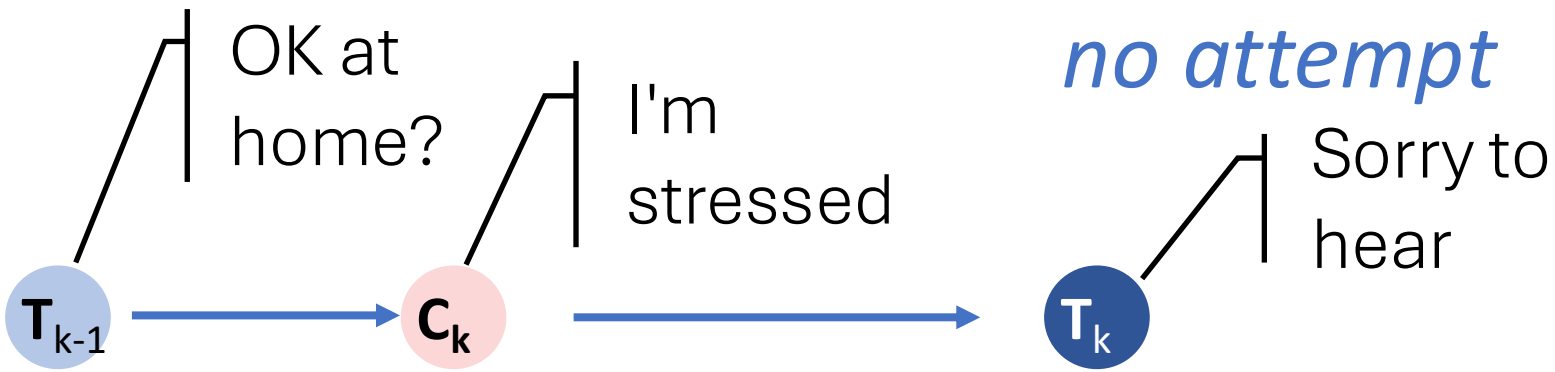
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



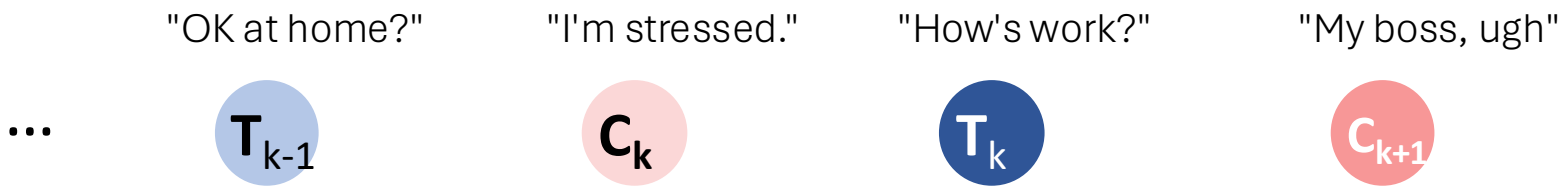
Intuition: (non)redirection patterns, in 2-D

T and C (Therapist/Client or Counselor/Texter)



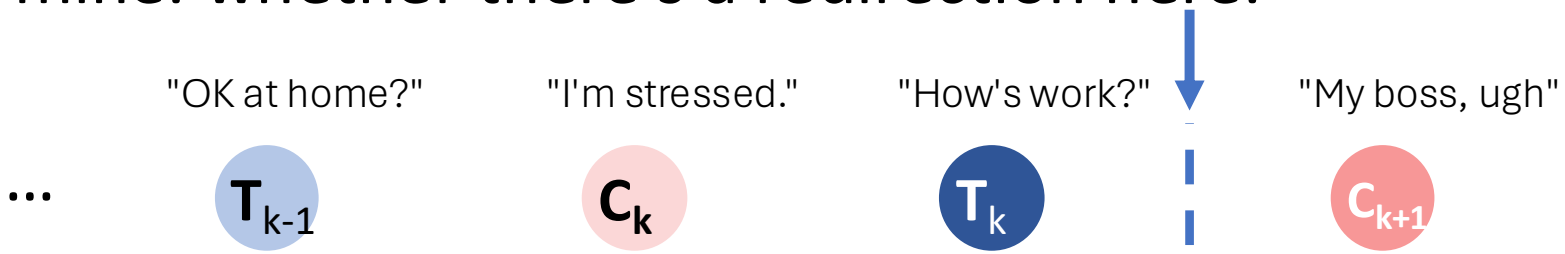
No redirection: no attempt

Formalizing redirection in real life = 1-D



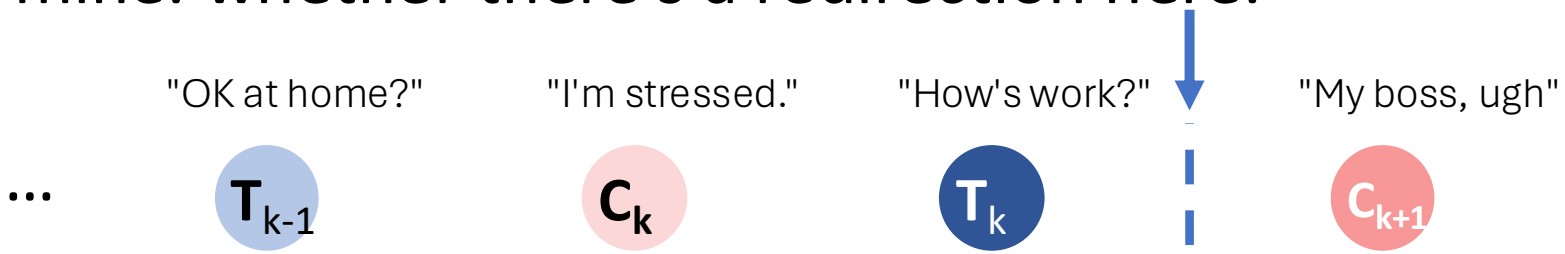
Formalizing redirection in real life = 1-D

To determine: whether there's a redirection here:



Formalizing redirection in real life = 1-D

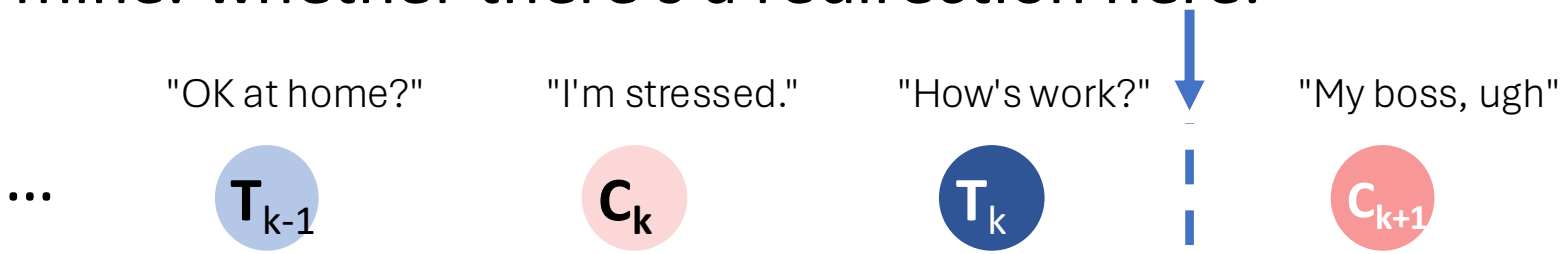
To determine: whether there's a redirection here:



Must check for both an attempt by T and an accept by C?

Formalizing redirection in real life = 1-D

To determine: whether there's a redirection here:



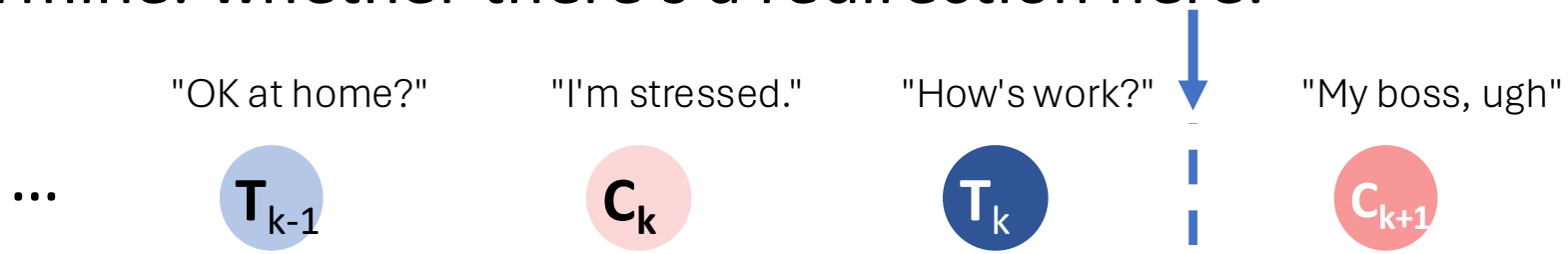
Must check for both an attempt by T and an accept by C?

Idea: compare against hypothetical:

Could C_{k+1} also follow something *else* T could *reasonably* have said?

Formalizing redirection in real life = 1-D

To determine: whether there's a redirection here:



Must check for both an attempt by T and an accept by C?

Idea: compare against hypothetical:

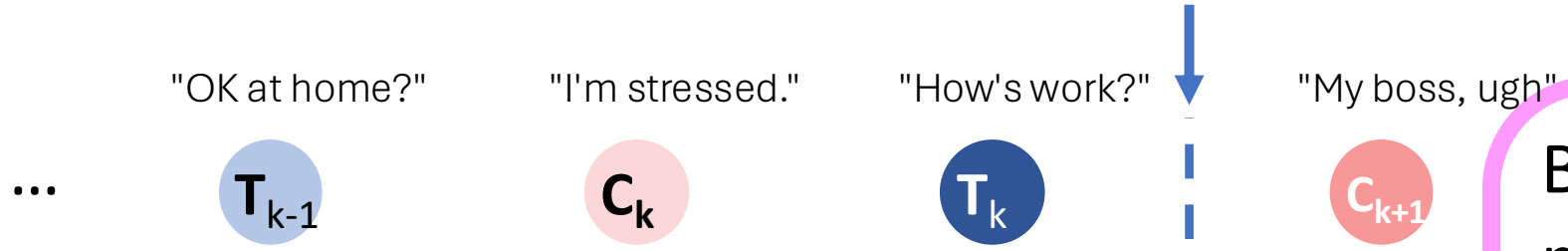
Could C_{k+1} also follow something *else* T could *reasonably* have said?

We choose T_{k-1} in particular.

"OK at home?"

Formalizing redirection in real life = 1-D

To determine: whether there's a redirection here:



Must check for both an attempt by T and an accept by C?

Idea: compare against hypothetical:

But repeating T_{k-1} might make the conversation weird …

So, ignore everything *before* C_k !

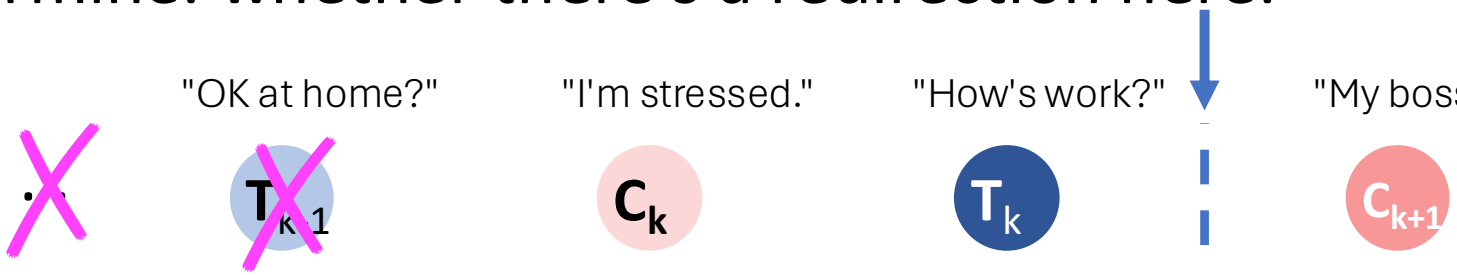
Could C_{k+1} also follow something *else* T could *reasonably* have said?

We choose T_{k-1} in particular.

"OK at home?"

Formalizing redirection in real life = 1-D

To determine: whether there's a redirection here:



Must check for both an attempt by T and an accept by C?

Idea: compare against hypothetical:

But repeating T_{k-1} might make the conversation weird …

So, ignore everything *before* C_k !

Could C_{k+1} also follow something *else* T could *reasonably* have said?

We choose T_{k-1} in particular.

"OK at home?"

Formalizing redirection (cont.)

Take the log-likelihood ratio between

Prob(C_{k+1} | C_k , T_k)

- "I'm stressed."
- "How's work?"
- "My boss, ugh."

and

Prob(C_{k+1} | C_k , T_{k-1})

- "I'm stressed."
- "OK at home?"
- "My boss, ugh."

Formalizing redirection (cont.)

Take the log-likelihood ratio between

Prob(C_{k+1} | C_k , T_k)

- "I'm stressed."
- "How's work?"
- "My boss, ugh."

and

Prob(C_{k+1} | C_k , T_{k-1})

- "I'm stressed."
- "OK at home?"
- "My boss, ugh."

If **>>**,
T_k differs from T_{k-1} (redirection *attempt*), **and**
C_{k+1} is better response to T_k (attempt *accepted*)

Formalizing redirection (cont.)

Take the log-likelihood ratio between

Prob(C_{k+1} | C_k , T_k) and **Prob**(C_{k+1} | C_k , T_{k-1})

- "I'm stressed."
- "How's work?"
- "My boss, ugh."

- "I'm stressed."
- "OK at home?"
- "My boss, ugh."

If **>>**,
T_k differs from T_{k-1} (redirection *attempt*), **and**
C_{k+1} is better response to T_k (attempt *accepted*)

If **<<** ,
T_k differs from T_{k-1} , **but**
C_{k+1} is better response for T_{k-1} (attempt *rejected*)

Meta-remark: on "unrefined" implementation

We want to see if the *core idea* has merit.

- We'll achieve interesting results even with *blunt approximations* and *lightweight models*, producing a *floor* on expected performance.

- Also, lightweight models can be run locally: important for data privacy

Prior measures: no joint action

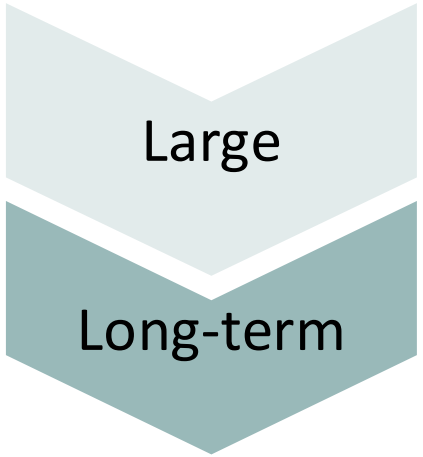
- Orientation [Zhang & D-N-M ‘20]
  - Captures only the attempt to shift the focus of the conversation.

- Similarity Difference
  - Doesn't consider attempt.
  - Compares similarities, not probabilities, but counterfactual is also "repetition"

- Uptake [Demszky, Liu, Mancenido, Cohen, Hill, Jurafsky, Hashimoto ‘21]
  - Doesn't consider attempt.
  - Does compare against counterfactual, but it's a random replacement.

Only *redirection* has a significant correlation with therapeutic outcome.

Therapy data: TalkSpace (access gained by long collaboration)



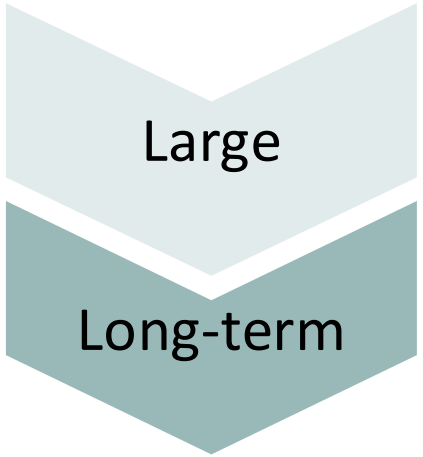
- 18K licensed therapists, 300K patients, 65M text msgs
- 26K therapies lasting over a year, 17K therapies have > 500 msgs

Probabilities via fine-tuned Gemma-2B w/ QLoRA on 8000 held-out therapies:

All computation must be *local* due to significant privacy concerns.

Slide design adapted from Vivian Nguyen

Therapy data: TalkSpace (access gained by long collaboration)



- 18K licensed therapists, 300K patients, 65M text msgs
- 26K therapies lasting over a year, 17K therapies have > 500 msgs

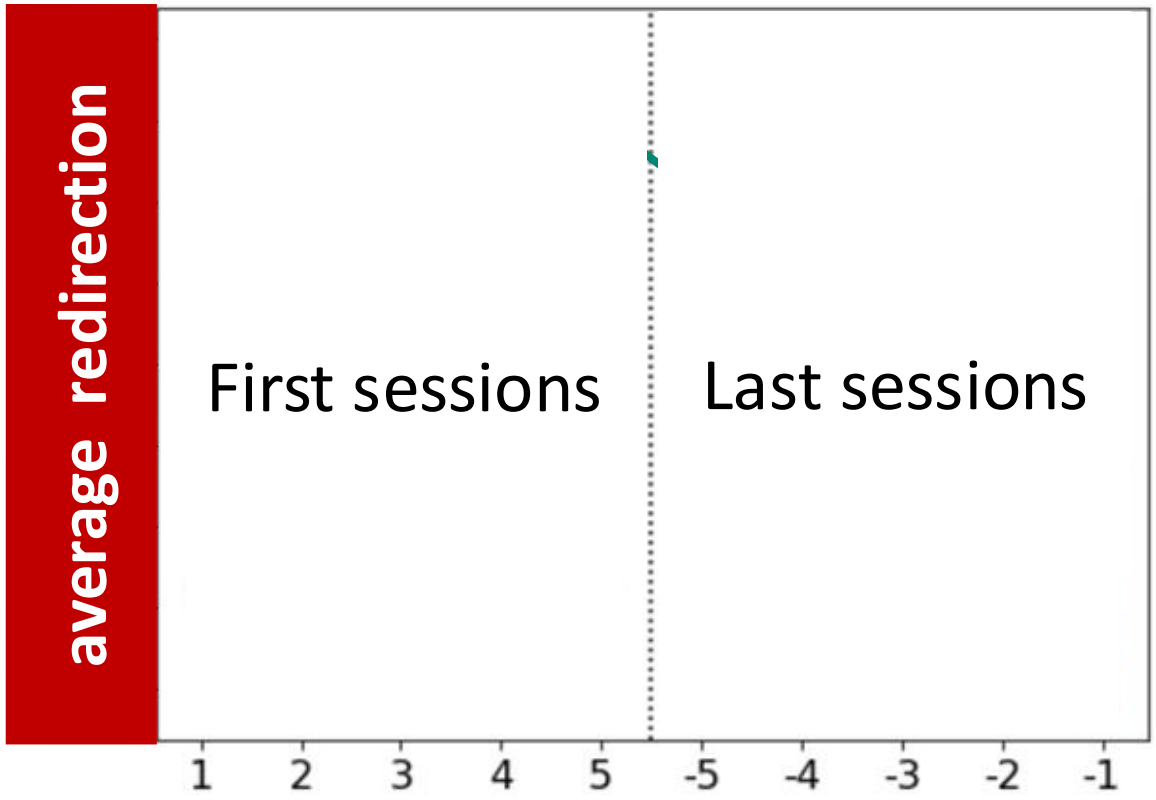
Probabilities via fine-tuned Gemma-2B w/ QLoRA on 8000 held-out therapies:

All computation must be *local* due to significant privacy concerns.

Notably: significant results achieved even with such a small model.

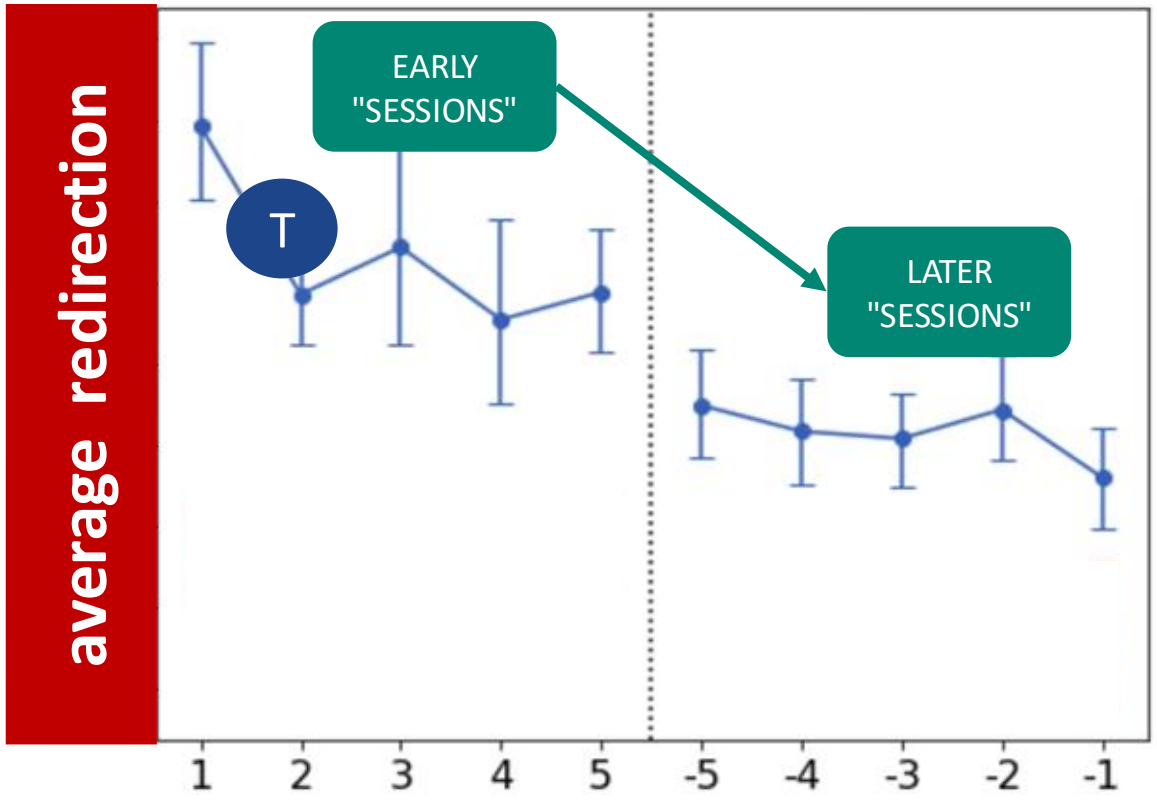
Intuition check: redirection and relationship evolution

Results on 3.7K therapies lasting at least 10 "sessions".
T = therapists, C = clients



Intuition check: redirection and relationship evolution

Results on 3.7K therapies lasting at least 10 "sessions".
T = therapists, C = clients



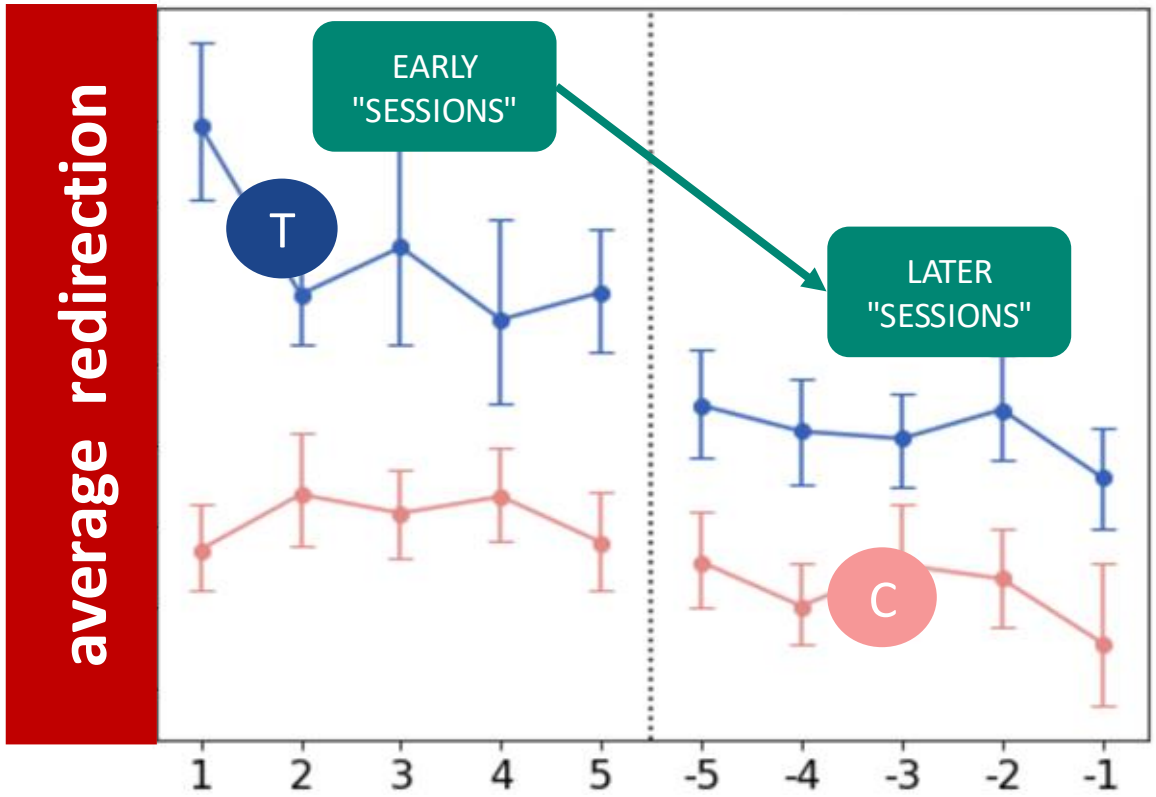
therapist

average redirection decreases.

($p < .0001$, Wilcoxon sign test)

Intuition check: redirection and relationship evolution

Results on 3.7K therapies lasting at least 10 "sessions".
T = therapists, C = clients

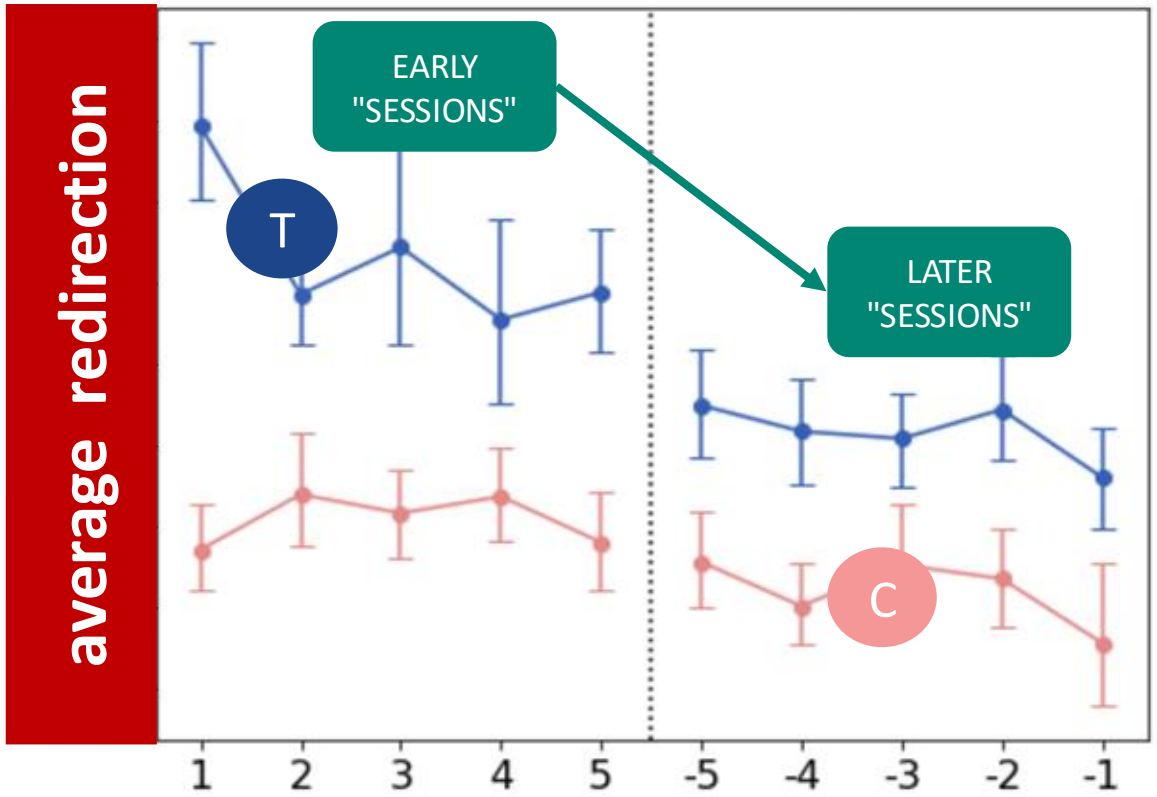


Both therapist and client's average redirection decreases.

($p < .0001$, Wilcoxon sign test)

Intuition check: redirection and relationship evolution

Results on 3.7K therapies lasting at least 10 "sessions".
T = therapists, C = clients



Both therapist and client's average redirection decreases.
($p < .0001$, Wilcoxon sign test)

Client's relative share increases.
($p < .0001$, Wilcoxon sign test)

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Unsuccessful relationships: client "ended" and cited problem w/ therapist, after at least 3 sessions (n=817)

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Unsuccessful relationships: client "ended" and cited problem w/ therapist, after at least 3 sessions (n=817)

Control relationships: no "end" from client, at least 3 sessions (n chosen to be 817)

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Unsuccessful relationships: client "ended" and cited problem w/ therapist, after at least 3 sessions (n=817)

Control relationships: no "end" from client, at least 3 sessions (n chosen to be 817)

In the 1st three sessions of unsuccessful relationships, clients redirect less (p-val .02, Mann-Whitney).

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Unsuccessful relationships: client "ended" and cited problem w/ therapist, after at least 3 sessions (n=817)

Control relationships: no "end" from client, at least 3 sessions (n chosen to be 817)

In the 1st three sessions of unsuccessful relationships, clients redirect less (p-val .02, Mann-Whitney).

Same attempt rate, so/but: therapists aren't accepting the attempts.

Does redirection correlate with relationship outcome?

- Patients can cancel / switch therapist (="end") at any time
- May give reason for "end" (not shown to therapist)

Unsuccessful relationships: client "ended" and cited problem w/ therapist, after at least 3 sessions (n=817)

Control relationships: no "end" from client, at least 3 sessions (n chosen to be 817)

In the 1st three sessions of unsuccessful relationships, clients redirect less (p-val .02, Mann-Whitney).

Same attempt rate, so/but: therapists aren't accepting the attempts.

No significance for therapist-initiated redirection, or the other prior measures.

Next, our second type of moment: pivotal ones

… when the outcome seems to "*hang in the balance*" depending on what is said next.

Visualizations by Moritz Stefaner, Dario Taraborelli, and Giovanni Luca Ciampaglia

Intuition example

Crisis Text Line conversation snippet, manually paraphrased for privacy.

Texter has just described panic-attack symptoms and suicidal thoughts.

I'm at the hospital right now screening for a surgery tmw

Would it be okay to share your name with me?

John. Why?

It's just to show you the respect you deserve,
and makes our conversation feel more personal.

I wish I had a friend like you

Intuition example

Crisis Text Line conversation snippet, manually paraphrased for privacy.

Texter has just described panic-attack symptoms and suicidal thoughts.

I'm at the hospital right now screening for a surgery tmw

[illegible]

Conversation outcome seems to *hang in the balance,* depending on next reply.

It's just to show you the respect you deserve,
and makes our conversation feel more personal.

I wish I had a friend like you

Domain: Crisis Text Line

24/7 crisis counseling service; goal is to guide texter to a calmer state.

Initial data: 1.5M conversations from 2015-2020.

Domain: Crisis Text Line

24/7 crisis counseling service; goal is to guide texter to a calmer state.

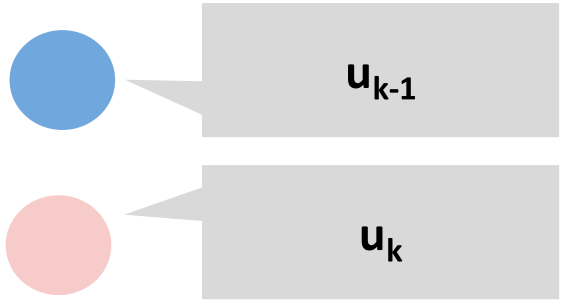
Initial data: 1.5M conversations from 2015-2020.

Unsuccessful interaction: *Abandonment* by the texter partway through

- Counselors react with a standardized protocol

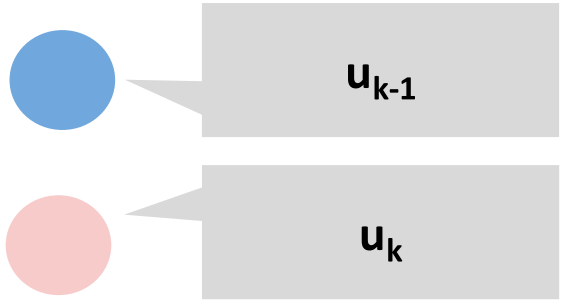
Successful interaction: texter responds to follow-up survey that the counselor was helpful.

Formalizing pivotalness

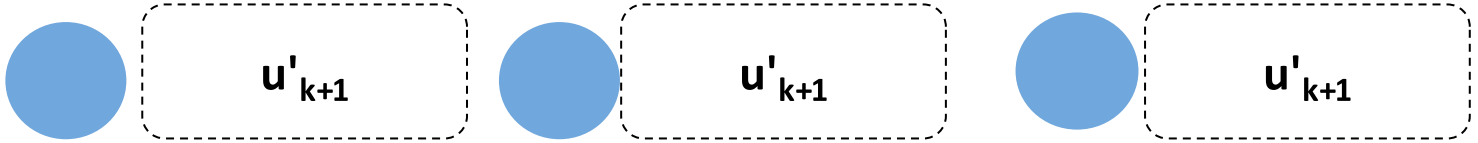


Slide design follows from Vivian Nguyen

Formalizing pivotalness

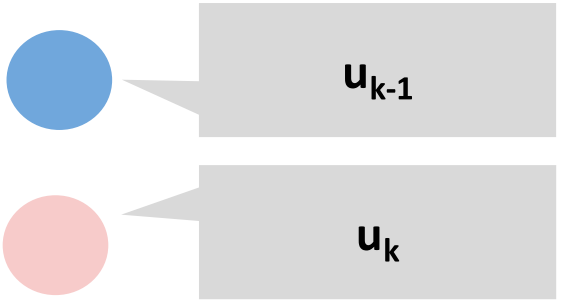


Potential immediate replies



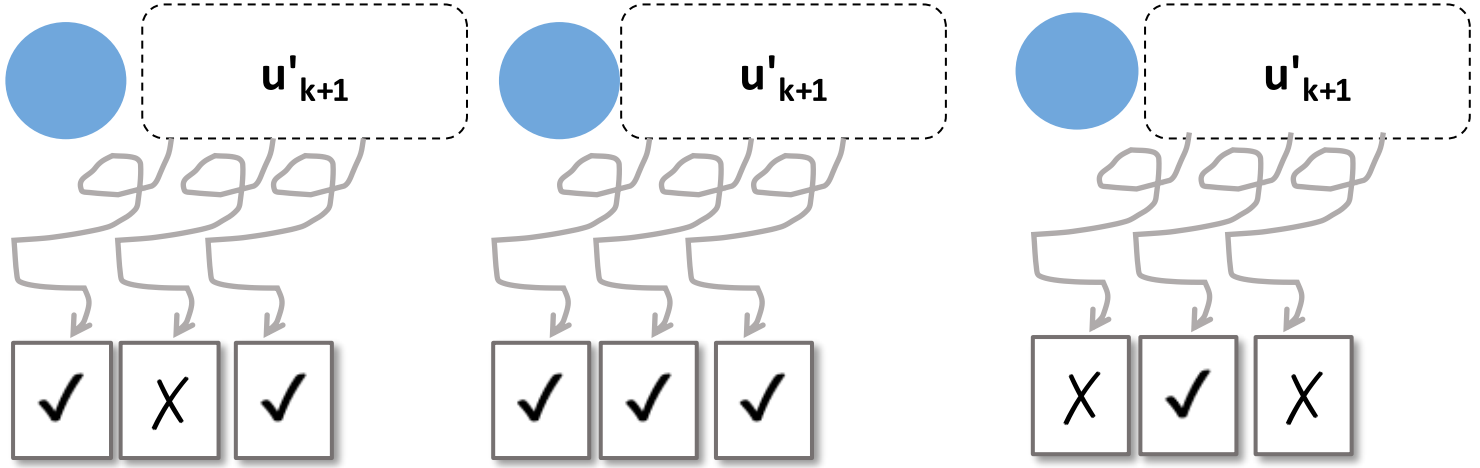
Slide design follows from Vivian Nguyen

Formalizing pivotalness



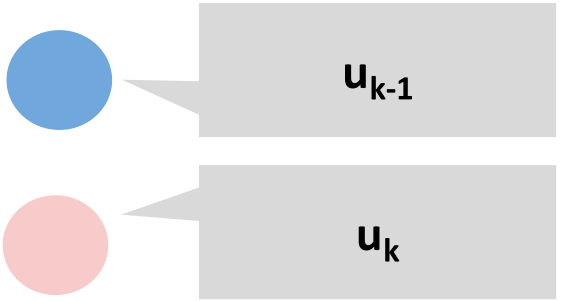
Potential immediate replies

Eventual outcome



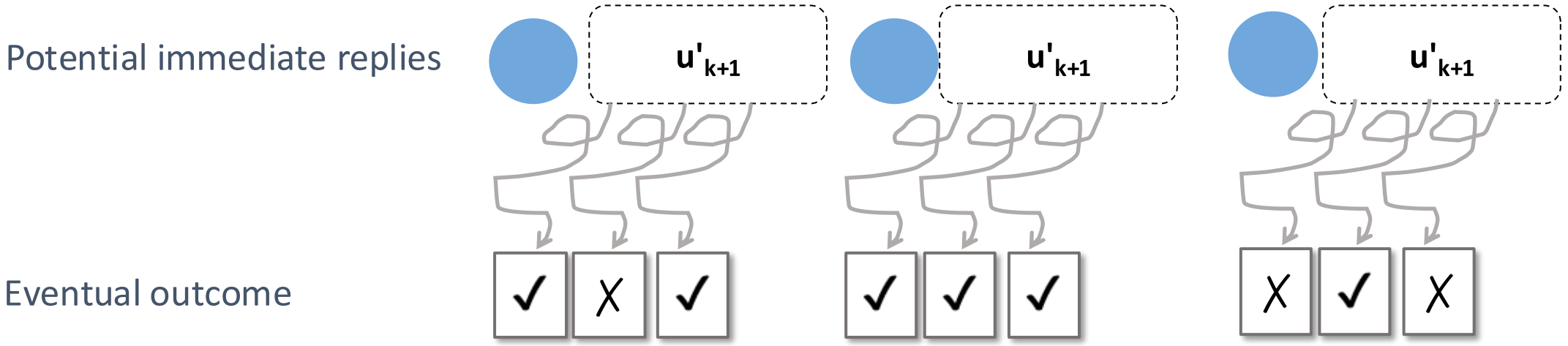
Slide design follows from Vivian Nguyen

Formalizing pivotalness

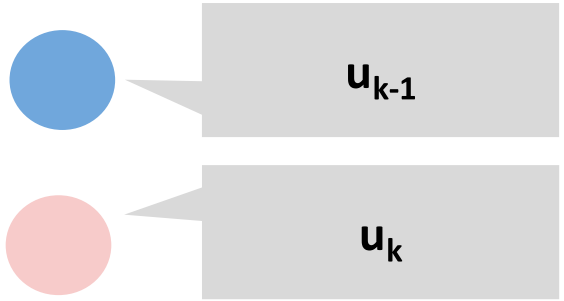


PIV: Variance over u'_{k+1} of $P(\text{success} \mid u_1 \ldots u_k\ u'_{k+1})$

This is an operationalization of *surprise* (Ely et al. 2015).

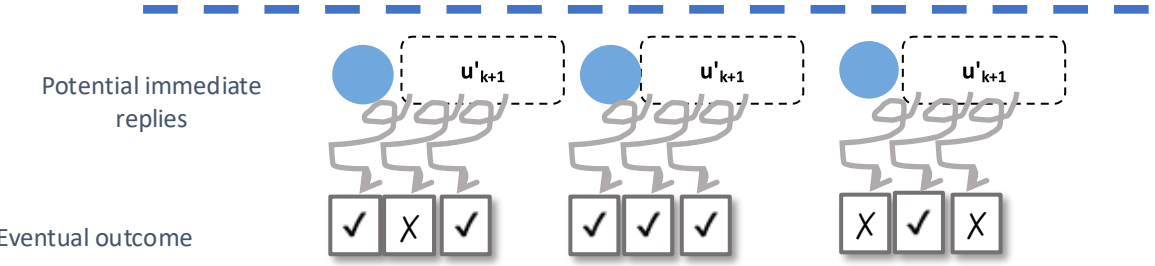


Formalizing pivotalness

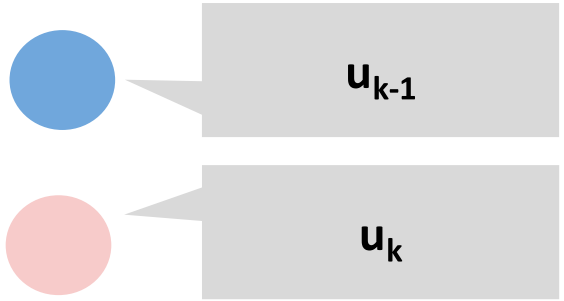


PIV: Variance over u'_{k+1} of $P(\text{success} \mid u_1 \ldots u_k\ u'_{k+1})$

This is an operationalization of *surprise* (Ely et al. 2015).



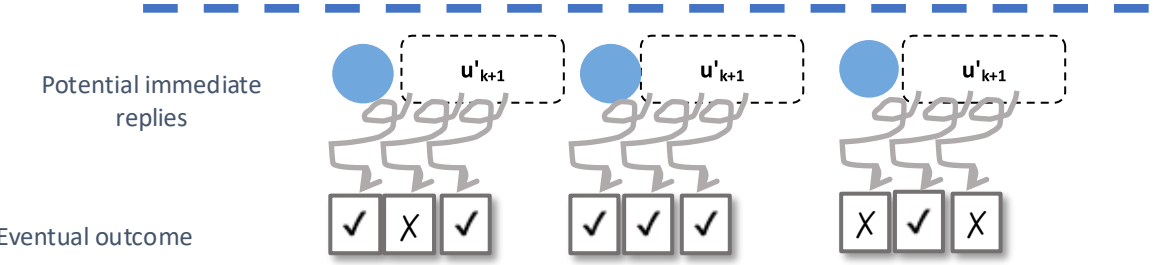
Formalizing pivotalness



PIV: Variance over u'_{k+1} of P(success | u_1 … u_k u'_{k+1})

This is an operationalization of *surprise* (Ely et al. 2015).

- Sample (just) the next reply u'_{k+1} via LLM, trained on heldout convs.
- Learn outcome *forecaster* for P(success | …) (Chang & DNM '19) on another balanced heldout set (with last 3 utterances removed, since they have obvious cues)



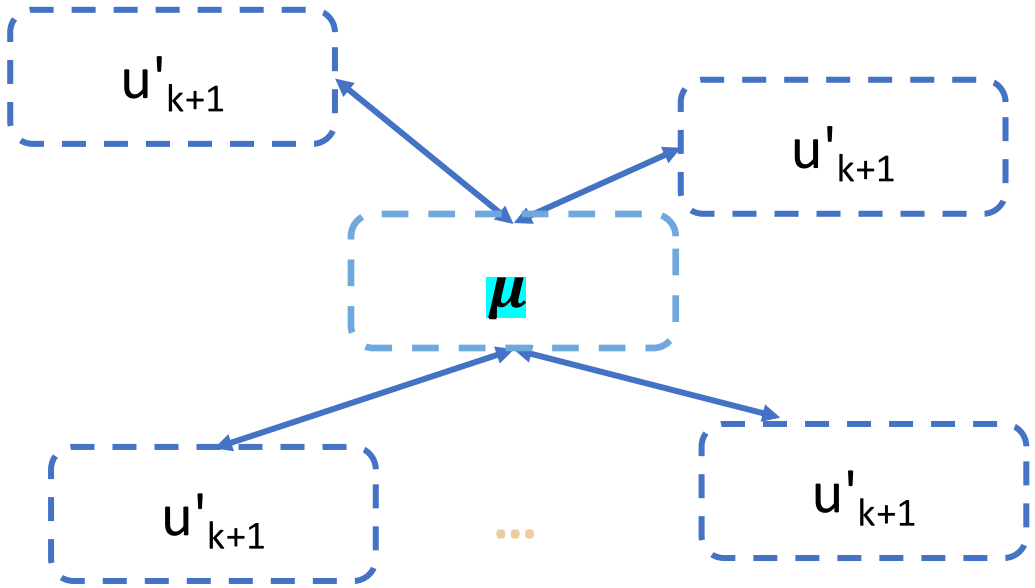
Slide design follows from Vivian Nguyen

Baseline measure: Range

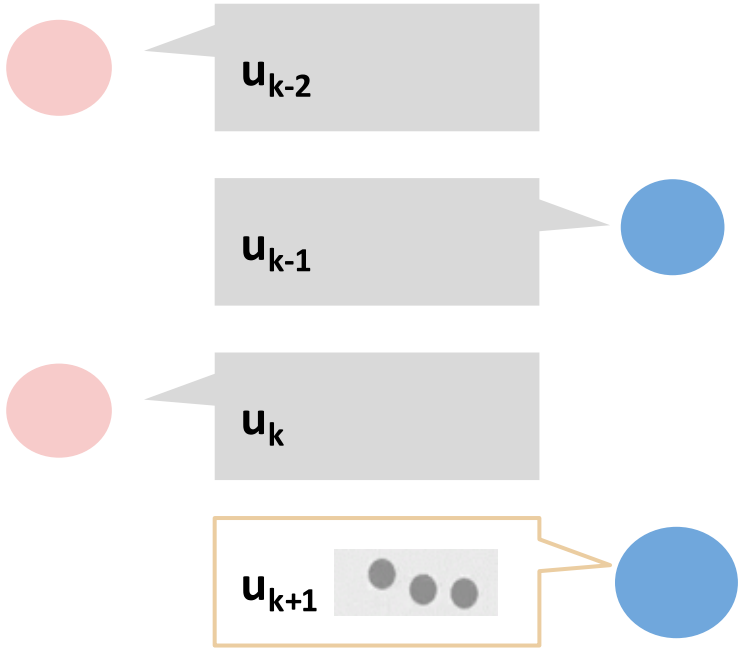
PIV: Variance over u'_{k+1} of P(success| $u_1 \dots u_k$ u'_{k+1})

Range: looks just at the *spread* of *next possible replies*

- Compute the mean μ of the sampled u'_{k+1}
- Compute average cosine distance between the sampled u'_{k+1} and μ

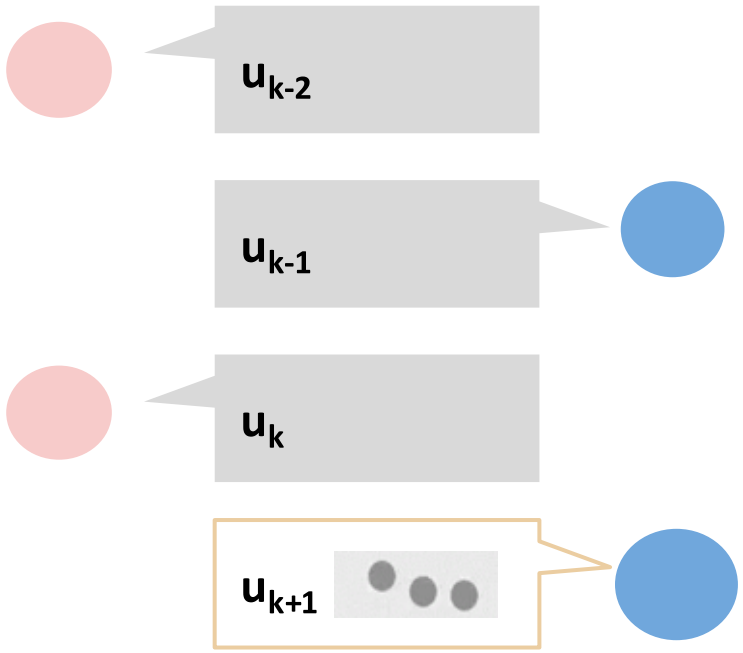


Intuition validation: response-time correlation



Data: 1000 conversations, balanced on outcome.

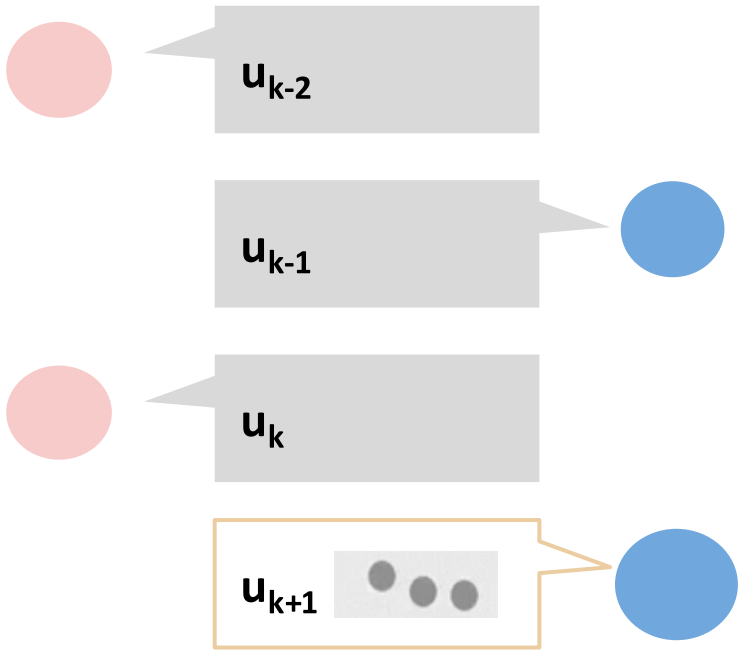
Intuition validation: response-time correlation



Data: 1000 conversations, balanced on outcome.

Counselors take longer to respond in high-PIV moments than low-PIV moments (p = 0.001, Mann-Whitney) ...

Intuition validation: response-time correlation



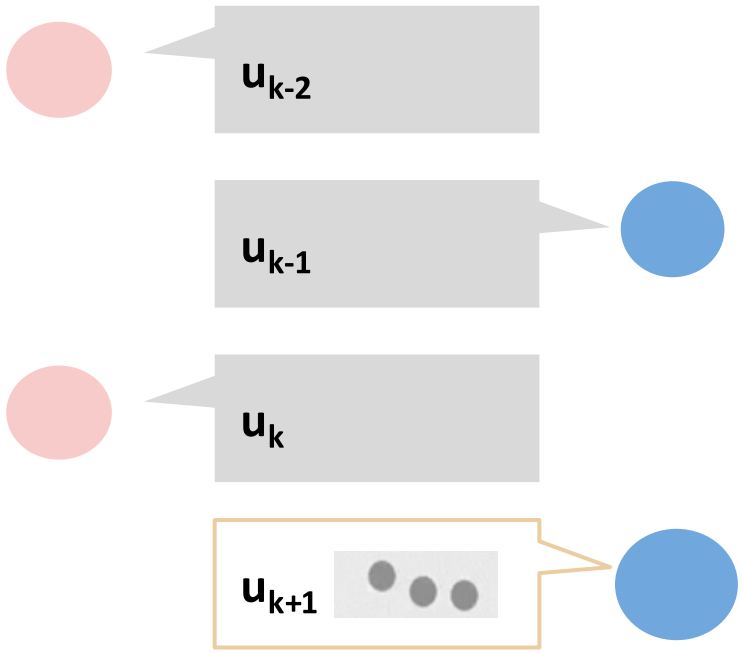
Data: 1000 conversations, balanced on outcome.

Counselors take longer to respond in high-PIV moments than low-PIV moments (p = 0.001, Mann-Whitney) ...

... despite there being no significant difference in reply length.

Slide design adapted from Vivian Nguyen

Intuition validation: response-time correlation



Data: 1000 conversations, balanced on outcome.

Counselors take longer to respond in high-PIV moments than low-PIV moments (p = 0.001, Mann-Whitney) ...

... despite there being no significant difference in reply length.

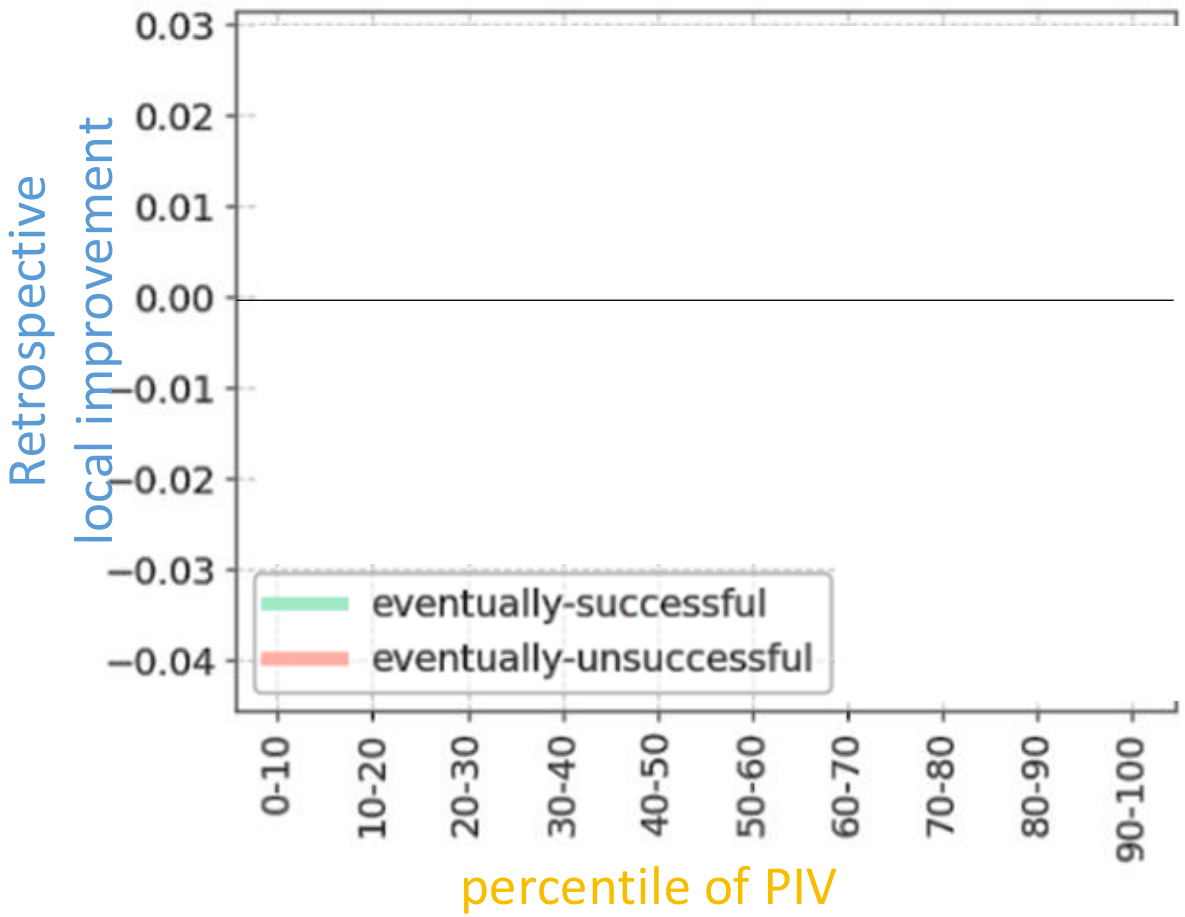
Range shows no such significant timing difference.

Slide design adapted from Vivian Nguyen

(How) Do actions in pivotal moments affect the outcome?

We can *retrospectively* estimate impact of the counselor's ***actual*** $\mathbf{u_{k+1}}$:

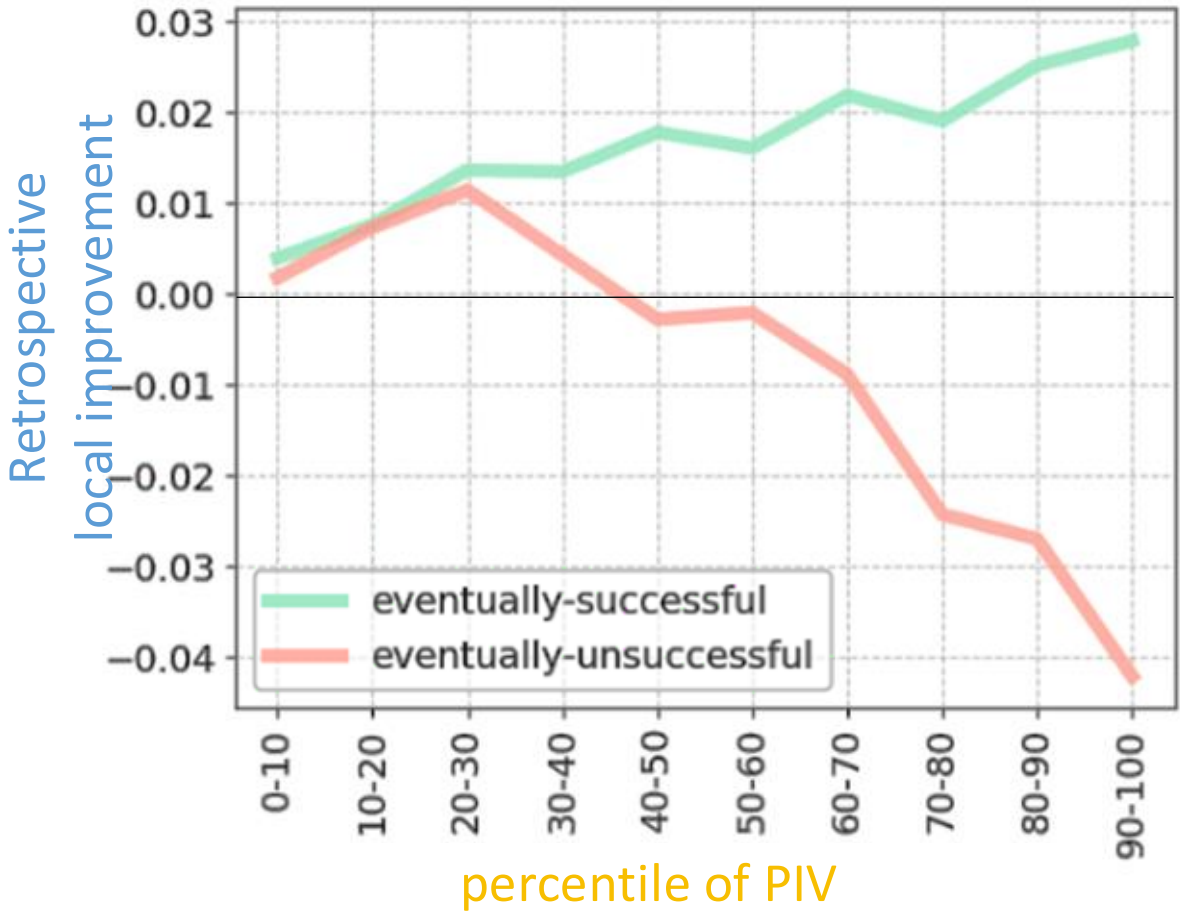
$$RI = (success | u_1 \dots u_k \ \mathbf{u_{k+1}}) - P(success | u_1 \dots u_k)$$



(How) Do actions in pivotal moments affect the outcome?

We can *retrospectively* estimate impact of the counselor's ***actual*** $\mathbf{u_{k+1}}$:

$$RI = (success | u_1 \dots u_k \ \mathbf{u_{k+1}}) - P(success | u_1 \dots u_k)$$



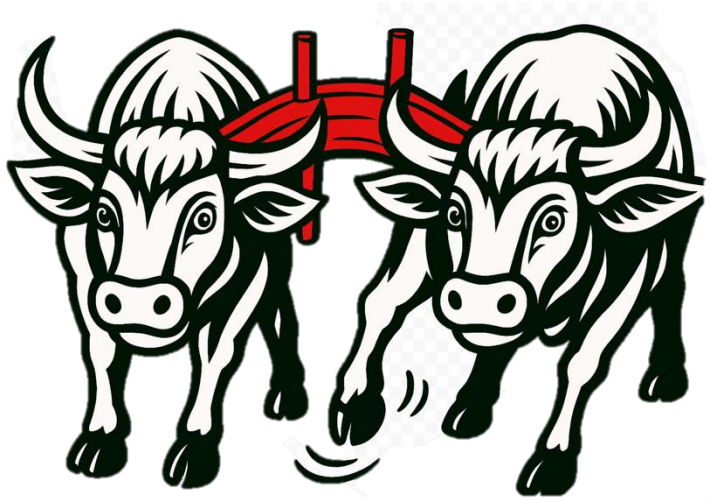
Low-PIV? u_{k+1} has little impact.

For eventually **successful** conversations, the largest changes are at the most pivotal moments, and are positive.

In **unsuccessful** ones: ditto, but the changes are negative.

Conclusion: we can detect "**key moments**" relevant to conversation **outcome!**

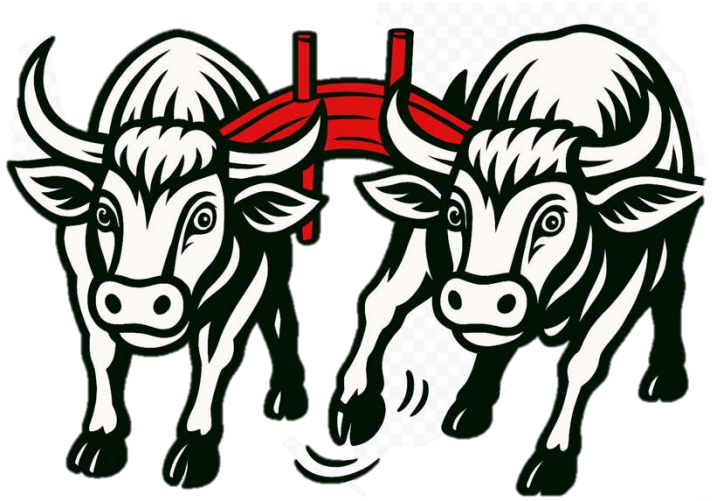
Redirection (joint)



Pivotal moments

Conclusion: we can detect "**key moments**" relevant to conversation **outcome!**

Redirection (joint)



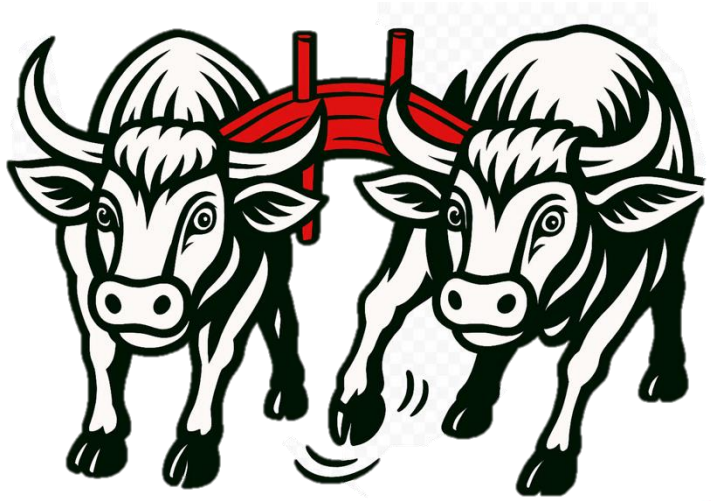
Pivotal moments

We wanted to see if the *core ideas* have merit.

We achieved interesting results even with *blunt approximations* and *lightweight models*, producing a *floor* on expected performance.

Conclusion: we can detect "**key moments**" relevant to conversation **outcome!**

Redirection (joint)



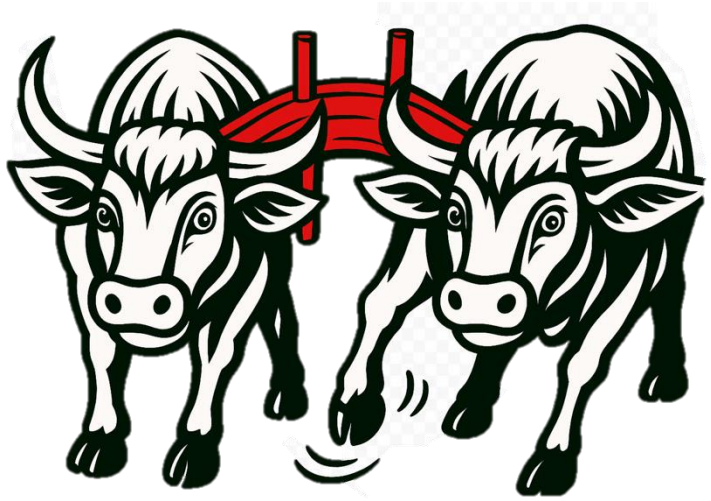
Pivotal moments

See improvements? Or applications to your domain? Fantastic! Code: https://convokit.cornell.edu/

Ex: GATech/Northwell Health used redirection to evaluate their CALM-IT system [Nguyen et al. '26]

Thanks, and let's see how *you* continue the conversation!

Redirection (joint)



Pivotal moments

See improvements? Or applications to your domain? Fantastic! Code: https://convokit.cornell.edu/

Ex: GATech/Northwell Health used redirection to evaluate their CALM-IT system [Nguyen et al. '26]