Randomized Controlled Trials as the Gold-Standard for Evaluating LLMs: A Primer for Biomedical NLP Researchers

Vicente Ivan Sanchez Carmona and Shanshan Jiang and Bin Dong
Ricoh Software Research Center (Beijing) Co., Ltd
{Vicente.Carmona, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com

Abstract

Large Language Models (LLMs) are no longer mere laboratory objects of study. LLMs have become everyday tools in society across diverse populations and domains. In clinical contexts, LLMs have already been devised as clinical support applications. However, along with benefits, negative or adverse effects might arise, such as LLMs potentially providing psychologically distressing advice to adolescents when used for mental health support. This raises questions on the benefits of LLMs and calls for real-world evaluations: Are LLMs really helpful and effective for the intended purposes people are using them or will use them for? To answer this type of question we propose to use Randomized Controlled Trials (RCTs). RCTs are considered the most strict experimental design in the fields of Medicine, Psychiatry, Psychology, among others; however, the use of RCTs in the NLP field is almost negligible. In spite of the NLP field being the de facto locus of research on LLMs, other fields, prominently Medicine, are leading the RCT evaluations on LLMs. In this primer paper, we present a concise introduction to the principles of RCTs to guide NLP researchers to design RCT studies for evaluating LLMs.

1 Introduction

In a recent submission to the journal of *Computational Linguistics*, Reiter (2025) argues that the ACL community places a small emphasis on real-world evaluations of NLP applications. While we applaud and echo this argument, we acknowledge that, to the detriment of both the ACL community and the final users, lacking real-world evaluations not only limits our understanding of the efficacy and utility of these applications on those users, it can lead to adverse scenarios, such as users ingesting chemicals with potentially toxic properties as advised by an LLM (Eichenberger et al., 2025).

Critically, evaluation deficits of NLP applications in clinical contexts may potentially open the possibility for negative scenarios such as adolescents being psychologically affected after interactions with LLMs, chatbots, or conversational agents since, according to a survey, it has been estimated that approximately one in eight adolescents and young adults have engaged with online LLMs for mental health advice and to alleviate mental distress (McBain et al., 2025). Although previous works, mainly from clinical fields, have already assessed the impact of generative systems on this outcome, the overall evidence is inconclusive (Feng et al., 2025). In addition, online LLMs are being adopted to other use cases. For example, students of different ages access LLMs for requesting help with homeworks or solving complex problems (Gasaymeh et al., 2025; Wang and Fan, 2025). However, in the NLP field there is scarce scientific evidence on how effective and useful are LLMs to address this type of clinically (and other real-world) relevant cases.

Benchmark evaluations of LLMs such as Big Bench (Srivastava et al., 2023) or MMLU (Hendrycks et al., 2021) may not be representative of real-world, naturalistic scenarios and are not suitable for quantifying the efficacy of LLMs in clinical use cases; these benchmarks assess how well LLMs solve problems, or answer questions, across several domains; but they cannot assess the usefulness of LLMs to allow users to achieve their goals: If an LLM is used as a support counselor by adolescents suffering anxiety, no benchmark evaluation can assess if the LLM was effective in reducing the adolescents' anxiety symptoms.

In this paper, we provide a primer for biomedical NLP researchers on what is considered the *gold-standard* method for evaluating new treatments in clinical disciplines, namely Randomized Controlled Trials (RCTs) (Torgerson and Torgerson, 2008; Kamp et al., 2026; Lilliengren, 2023). RCTs, a type of clinical trial (Hackshaw, 2009), are the gold-standard due to their rigorous experimental

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 392–406

design which allows researchers to control for biases and confounders that threaten to distort the true effect of the target treatment; thus, RCTs can lead to an unbiased and valid conclusion of the efficacy of a treatment (Bishop and Thompson, 2023; Hackshaw, 2009). We firmly believe that adopting RCTs will allow NLP researchers to stringently evaluate LLMs on real-world cases, and also to foster interdisciplinary research with clinical and healthcare communities. We truly hope this primer on RCTs will be widely used as a reference and guide in biomedical NLP in order to keep pushing the scientific progress of the field.

2 Background and Previous Works

2.1 Background

We start our account of RCTs by providing basic terminology. An RCT is an experiment, or study, conducted to provide a valid inference: Is a new treatment useful or effective for some goal when applied to participants? (Jadad and Enkin, 2007) A *treatment* is an intervention, a program, or in general, a mechanism which is hypothesized to provide a benefit to a group of people–the *participants* of the experiment. Similar to experiments in NLP where researchers test if adding a new module to an LLM helps the LLM to improve accuracy scores on a dataset, we seek to evaluate if our proposed treatment–for example an LLM–helps participants to attain a goal, such as distressed adolescents reducing their depression symptoms. In other words, we want to estimate the *effect* of the treatment, i.e. we want to quantify how much our treatment helped participants to attain a goal as measured by an outcome measure, such as a psychological questionnaire. This follows the same spirit as in NLP research where researchers estimate the effect of an improved or new model by comparing its mean accuracy score with that from a baseline model. However, when the evaluation of a treatment includes people, several *biases* and *confounders*–spurious factors–may be present which can distort the treatment effect threatening what is called the *internal validity* of the experiment, i.e. the capacity of a study to provide valid results. To counteract these spurious factors, researchers add another group of participants to the experiment called the *control group*–the equivalent of a baseline model in NLP–which will be compared to the *treatment group* (the group of participants on which we test our treatment), thus rendering valid results of the treatment's efficacy. In the following sections, we will provide the basic scientific principles of RCTs while showing how to avoid biases and confounders to design valid RCTs to evaluate LLMs on real-world scenarios.

2.2 RCTs for Evaluating Treatments in Clinical Disciplines

RCTs are extensively used across clinical disciplines. For example, in Psychiatry, RCTs have been used to evaluate treatments to improve the well-being of autistic children, such as fostering early social engagement (Carruthers et al., 2024), improving their socio-emotional and sleeping patterns (Boone et al., 2022), and developing their interaction and communication skills (Green et al., 2022). In Medicine, researchers have used RCTs to evaluate exercise programs for rehabilitation of people who have suffered a stroke (Moore et al., 2016), to assess interventions for promoting wound healing (Schneider et al., 2025), and to provide reliable evidence of the quality of new methods of cancer surgery (Ceelen and Soreide, 2023) and bone surgery (Jacobsen et al., 2020). Without RCTs assessing these type of critical treatments, end-users could be at risk of facing adverse consequences (Torgerson and Torgerson, 2008).

2.3 RCTs for Evaluating LLMs in Clinical Scenarios

Recently, RCTs have also been used to evaluate the efficacy of LLMs. For example, an LLM-based chatbot demonstrated comparable efficacy to a nurse hotline in reducing anxiety and depression symptoms of participants (Chen et al., 2025). Also, an RCT showed the usefulness of ChatGPT for physicians and medical students to conduct research but not any benefit for clinical decision-making (Weuthen et al., 2025). Furthermore, two RCTs showed that ChatGPT's role as a learning tool enhanced the performance of both dental and medical students for the lectures of clinical operative skills and orthopedics, respectively, compared to control groups (Huang et al., 2025; Gan et al., 2024). Remarkably, all of these evaluations come from medical venues. We strongly advocate for biomedical NLP researchers to be also actively engaged in these evaluations using this paper as a primer guide for learning the scientific principles for designing RCTs for LLMs.

3 Hypothetical Studies: Exemplifying the Evaluation of LLMs

To exemplify some points throughout the paper, we propose two hypothetical studies based on AI-enabled clinical support applications devised by the American Psychological Association (Association, 2025): 1) evaluating the efficacy of a tutor LLM for medical students aimed at improving their training in clinical diagnosis; 2) evaluating the efficacy of a complementary counseling LLM for mental health support aimed at reducing symptoms of distressed adolescents suffering from a psychological condition, such as depression or anxiety.

4 The Perils of Before-and-After Studies

It may seem straightforward to evaluate the efficacy of a new treatment by applying it on a group of participants (a treatment group) and measuring how much they benefited from the treatment based on an outcome measure. This study type is called a *before-and-after* or *pre-test-and-post-test* study (Ho et al., 2018; Sanders, 2019). Let us use our hypothetical example of a tutor LLM: We form one clinical diagnosis class in a school to apply the LLM as a treatment; we hypothesize that the LLM will cause this group to improve their test scores at the end of the study. Then, we assess our hypothesis by first obtaining baseline test scores–the *pre-test* scores–before the LLM is applied; then, we apply the LLM to the group to tutor them on diagnostic problems for a period of time; finally, at the end of the study we re-test the students to obtain their final scores–the *post-test* scores. One way to assess that this group benefited from the tutor LLM is by calculating how much their post-test scores increased with respect to the pre-test scores (Sanders, 2019): If the mean post-test score is higher than the mean pre-test score then we conclude that the LLM had a positive effect by helping the students to achieve such an improvement. However, this conclusion might be plagued with biases and confounders possibly rendering it invalid. These spurious factors could distort the test scores and we cannot disentangle their effects from the LLM's effect in this type of study (Moseley and Pinheiro, 2022). The solution to this problem is using an RCT. Nevertheless, a before-and-after study forms the basis of RCTs and we must learn which biases and confounders we must be cautious of.

4.1 Biases and Confounders

Biases are factors which systematically introduce an error in the form of a deviation to the treatment effect, but cannot be measured in specific units (Ho et al., 2018; Bishop and Thompson, 2023). On the other hand, confounders are measurable variables (unrelated to the treatment) which effect can be conflated with the effect of the treatment (Kingsley and Robertson, 2020) giving us the false impression that the treatment is (or not) useful. In before-and-after studies, confounders can usually be controlled–i.e. their effect can be neutralized; however, effects from biases can only be reduced–only RCTs can neutralize or eliminate them. Below we show the most relevant biases researchers must be aware of before any study is conducted.[1]

Placebo bias Participants volunteering for a new treatment mainly do so due to their willingness to recover from a problematic situation or condition. Evidence in the literature points to this willingness having a psychological effect on the outcomes of the participants regardless of receiving an effective or ineffective treatment; i.e. participants exhibit a placebo effect (Beins and McCarthy, 2018). This effect's magnitude could be so pronounced as to spuriously boost the treatment effect; for example, the adolescents in our hypothetical study may elicit a placebo effect after knowing that they will take part in an experimental group where the supporting counselor is a state-of-the-art LLM; this effect is elicited in response to their high expectation or enthusiasm in the LLM to help them alleviate their depression symptoms, effectively leading to a reduction in these symptoms, as it has been seen before in other cases (Bishop and Thompson, 2023; Torgerson and Torgerson, 2008). We note that the main mechanism of the placebo bias is still not fully understood (Piantadosi, 2017), but in our view, a possible explanation of this bias is that the participants mindset changes in a positive way, due to their high enthusiasm or motivation in receiving a novel treatment, aligning to the direction of the treatment effect; i.e. the participants hold the belief that the treatment will work, and they

[1] Other biases affecting the treatment that can be controlled via RCTs are the *Hawthorne* effect (Torgerson and Torgerson, 2008), the *maturation* bias (Marsden and Torgerson, 2012), the *demand characteristics* bias (Beins and McCarthy, 2018), the *regression to the mean* effect (Smith, 2015), the *history* bias (Marsden and Torgerson, 2012), the *experimenter* bias (Torgerson and Torgerson, 2008), or the *measurement* bias (Creswell and Guetterman, 2018). See Section A.1.

unconsciously exert a positive outcome by taking actions in parallel to the LLM's intervention such as changing daily-life habits or requesting extra-help outside the study, which will have a positive effect on the post-test scores regardless of whether the treatment really worked or not. This bias cannot be controlled in before-and-after studies. The optimal control is via an RCT (Section 6.2).

Test sensitivity or memorization bias Ideally, every participant in a before-and-after study takes two tests, a pre-test and a post-test, to assess their progress; however, this may induce a bias on the participants (Ho et al., 2018). For example, they may become sensitive to the pre-test in the sense of discovering what type of information or problems are used for evaluation; in this way, the participants may actively seek to learn any material aligned to that type of information or problems so as to intentionally improve the post-test evaluation giving the false impression that improvement in performance is due to the LLM when, in fact, it is due to the participants' sensitization and extra-effort (Marsden and Torgerson, 2012). Moreover, if the format or structure of the post-test is highly similar to that from the pre-test, then participants may get a better score just due to being familiarized with the test format or structure (Bishop and Thompson, 2023). This bias not only occurs in educational settings; it can also happen in clinical and psychological studies where participants may receive the same well-being survey or questionnaire as both pre-test and post-test–such as the Patient Health Questionnaire-9 (Yu et al., 2012) developed to assess depressive symptoms–in this case, participants may respond to the post-test questionnaire in a very similar way to how they responded to the pre-test questionnaire due to remembering their previous responses and unconsciously aligning to them (Navarro and Siegel, 2018). The best solution for this bias type is the RCT shown in Section 6.3.

Selection bias Another major threat to the internal validity of before-and-after studies is the problem of a biased sample of participants (Torgerson and Torgerson, 2008). The participants selected to the study may be biased towards a characteristic that could inflate or dilute the effect of the treatment. For example, if the recruitment is open to anyone wishing to take part in it, as is common in several studies, this will lead to one of the most common forms of selection bias: Recruiting people who are particularly enthusiastic to participate in a study mainly because they believe the new treatment can be a great help to their underlying problem or condition. Thus, this enthusiasm is a characteristic that has the potential to spuriously inflate the treatment effect via different mechanisms such as eliciting a placebo bias. In medical or psychological studies, other form of selection bias can take place: Participants willing to take part in a study may exhibit a degraded physical or psychological well-being which, if not profoundly serious, tends to naturally recover after given enough time regardless of receiving or not any treatment (Torgerson and Torgerson, 2008). These are just a few forms of selection bias; depending on the specific research problem many other forms can plunge into the study. RCTs are suited to control this bias.

Confounders effect Characteristics of the participants such as socioeconomic status, age, gender, etc. may have an impact on the post-test scores (Bishop and Thompson, 2023). For example, in our hypothetical study of a tutor LLM, recruiting only students from a high socioeconomic status risks confounding since access to high-cost curricular resources may influence post-test scores, distorting the LLM's effect. Similarly, in our hypothetical study of a counseling LLM, age effects can be confounded with the LLM's effect since younger (or older) participants may tend to recover more easily depending on the target condition. Hence, it is important to consider possible confounders, according to the specific problem, and control them, for example, via variation in the levels of the confounder, such as recruiting participants across varying ages, or via an RCT study.

5 RCTs: The Gold-Standard for Evaluations of LLMs

Since RCTs build on top of before-and-after studies, in this section we elaborate on their basic principles which allow for a strict control of the spurious factors occurring in before-and-after studies (shown in Section 4.1) and on the specific biases RCTs face as well as ways to control them.

5.1 RCTs Can Control Spurious Effects from Before-and-After Studies

The effects of biases and confounders in Section 4.1, which threaten the internal validity[2] of any before-and-after study, can be neutralized by

[2]See Section A.2 for a brief explanation of internal validity.

adding two key elements to this type of study: 1) a control group and 2) the randomized assignment of participants to control and treatment groups, where the treatment group is similar to the group in a before-and-after study; these elements constitute the basic and core structural features of a Randomized Controlled Trial (Creswell and Guetterman, 2018). In the following sections we elaborate on these characteristics.

5.2 Random Assignment of Participants to Groups: Controlling Biases and Confounders

The effects of biases and confounders can be neutralized, or counteracted, via the random allocation of participants to the control and treatment groups (Creswell and Guetterman, 2018). The goal with this action is to have the treatment and control groups nearly identical to each other in all participants characteristics and factors–including confounders and biases–except for the treatment to be evaluated; thus, the only difference between these two groups will be the effect of the treatment allowing us to easily estimate it. Simple randomization is a common method to achieve this as explained below.

Random assignment of participants to treatment and control groups is essential to balance the groups across observable and unobservable characteristics and factors that otherwise can lead to an invalid result. The simplest random method of allocation recommended in the literature is *tossing a coin* (Torgerson and Torgerson, 2008) where for each participant we obtain a computer-generated random binary label ("treatment" or "control"), or alternatively, we generate a random list of numbers, assign participants to numbers, and allocate odd numbers to one group and even numbers to the other. In any scenario, crucially, each participant must have the same chance to be assigned to any group. However, due to chance a confounder may not be fully balanced between control and treatment groups; to purposefully control this confounder we can resort to the method of *matched randomization* (see Section A.3).

5.3 RCTs Can Be Susceptible to Other Biases

RCTs may suffer from particular biases which are important to consider before conducting any study in order to propose any possible solution if required, as we show below (and in Section A.4).

Diffusion of treatment bias This bias, also known as *spillover effect* (Ariel et al., 2021), happens when participants in the treatment group share information of the treatment with participants in the control group in such a way that control participants can follow or adapt this information to their benefit having an effect on the post-test scores and thus threatening the study's validity (Creswell and Guetterman, 2018). A feasible policy to control this bias is to implement a double-blind setup, where neither the participant nor the researcher knows which participant is in which group (Bishop and Thompson, 2023; Friedman et al., 2015), in addition to a non-disclosure agreement where participants are requested to not sharing information from the study other than to relevant parties (such as the participants' parents if the participants are minors).

Resentful demoralization bias When participants assigned to the control group realize that they receive no treatment (or receive a traditional treatment which they believe is not effective), they may develop a feeling of resentment, demotivation, or demoralization which may have one or more consequences (Torgerson and Torgerson, 2008). One possible consequence is that control participants have a lower performance during the study than they usually do in normal days; this will have a negative effect on the post-test scores leading to artificially inflating the treatment effect. Another consequence is that control participants may seek for alternative treatments out of the RCT study; in our hypothetical example, students may seek help from a professor or engage with publicly available LLMs, such as ChatGPT, possibly leading to a spurious boost of their post-test scores. An additional consequence is that participants may quit the experiment due to a high impact of the demoralization or after rationalizing that their participation in the experiment will not lead to any benefit for them. A last possible consequence is that control participants may work harder during the study than they normally do due to their resentment of not being assigned any treatment (Creswell and Guetterman, 2018); this action will lead to inflated post-test scores which will minimize the true treatment effect. The resentful demoralization bias can be ameliorated or resolved either by a) giving a traditional treatment to the control group (such as a human tutor in our example) that can be regarded as an effective treatment (Beins and McCarthy, 2018), b)

reassuring control participants that they will have the same treatment as the treatment group after the experiment is finished (Bishop and Thompson, 2023), or c) replacing the control group with a placebo group (Torgerson and Torgerson, 2008) as we will elaborate on in Section 6.2.

6 Three Recommended RCT Designs to Evaluate LLMs

RCTs' experimental design can be adjusted to specific research questions, problems, and to analyze specific biases and confounders. We selected three main design types that we believe can be useful for biomedical NLP researchers to evaluate the efficacy of LLMs in real-world scenarios, from the basic *treatment vs. control* RCT design to more complex designs that aim to study the effect of relevant biases, namely the placebo bias and the test sensitivity or memorization bias which, we believe, can have a big impact on the outcome measure whenever studying the efficacy of LLMs across several domains, specially in educational and clinical contexts.

6.1 The Basic RCT Design: Treatment vs. Control Groups

A control group is most necessary to counteract the effects of several biases and confounders present in the treatment group via random allocation of the participants to these groups (Section 5.2). This design is the easiest and less costly to implement since only two groups of participants are required, where the suggested number of participants in each group is at least 50 in order to provide a precise estimate of the treatment effect (Navarro and Siegel, 2018). As mentioned before, only the treatment group receives the novel treatment and the control group receives either no-treatment or a traditional treatment (Beins and McCarthy, 2018). However, this design may not render the most precise treatment effect as the control group may not neutralize one possible bias, namely the placebo bias, since it tends to occur on participants who receive a novel treatment as it elicits the participants' high expectations that the treatment will be useful as imagined, overstating the treatment's effect (Piantadosi, 2017). This bias may or may not occur in an RCT; however, there is compelling evidence across fields, particularly Medicine and Psychology (Beins and McCarthy, 2018; Piantadosi, 2017), pointing to its frequent occurrence. To optimally control this bias it is suggested to add a placebo group.

6.2 Controlling the Placebo Bias: The Placebo Group

We introduced the placebo bias in Section 4.1. An optimal way to control this bias is to add a placebo group to the basic RCT design (or alternatively, to replace the control group with a placebo group) (Piantadosi, 2017). This group will receive a placebo[3] instead of a true treatment. The form of the placebo will depend on the form of the treatment under study; crucially, no placebo should provide any real treatment. For example, a placebo counterpart of a clinical diagnostics tutor LLM could be an LLM-based chatbot, in the style of the chatbot Eliza (Weizenbaum, 1966), that provides no support in diagnostics (and crucially, does not generate hallucinations), but responds to queries in a thoughtful-questioning way by asking the students what they think is a possible solution to a clinical problem, or why they think a candidate solution could be a good solution (or not).[4]

By adding the placebo group to our basic RCT design we now have the capability to not only remove the placebo effect from the treatment effect, but also to quantify it (Bishop and Thompson, 2023). More concretely, if we only desire to remove the placebo effect from the treatment effect, then we only need to subtract the mean post-test score of the placebo group from the mean post-test score of the treatment group since these two groups only differ in the treatment provided (novel treatment vs. placebo);[5] however, the placebo effect is mixed with all the other bias effects and we cannot quantify it. If we seek to quantify the placebo effect to develop an understanding of the strength of

[3] In simple and basic terms, a placebo is a mechanism mimicking or emulating a novel treatment, which also elicits high expectations from participants, but has no effect on them.

[4] We remark the potential complexity of proposing and implementing effective placebo counterparts of LLMs due to the inherent necessity of participants to interact with this placebo (e.g. via dialogues): It is difficult to mimic all the form or structure of the treatment without providing any functionality that could become in itself a treatment, while at the same time both not providing any negative effect and not allowing the participants to discover the status of the placebo (specially in long-term studies). For a deeper consideration of this issue and to help identifying an appropriate approach tailored to the needs of a specific study, we suggest previous works in the literature of Education (Hyldgaard, 2020; Adair et al., 1990; Aycock et al., 2018) and Psychology (Blease, 2018; Kirsch et al., 2016; Enck and Zipfel, 2019; Popp and Schneider, 2015; Gaab et al., 2019).

[5] This also removes the effects of all other biases and confounders explained in Section 4.1.

the participants' expectations regarding the efficacy of LLMs in real-world cases, then we simply subtract the mean post-test score of the control group from the mean post-test score of the placebo group–assuming that these two groups only differ in the placebo effect due to the optimal random allocation of participants to the three groups; nonetheless, for a more precise estimation, the recommended method is ANCOVA (Section 7.2).

6.3 The Solomon Design

This RCT design is considered as the ideal design to better quantify and understand the effect of the pre-test on the post-test scores (Solomon, 1949). This, in turn, allows us to tackle the test sensitivity or memorization bias (Section 4.1).

The Solomon design comprises 4 groups; it is a direct extension of the basic RCT study (Section 6.1) to which we add two more groups: Another treatment group receiving the exact same treatment, and another control group receiving either the same traditional treatment or non-treatment. However, these two groups will not receive a pre-test. Let us label our four groups where T_1 and T_2 are the two treatments groups, T_1 receiving a pre-test but not T_2, and C_1 and C_2 are the two control groups where C_1 receives a pre-test but not C_2. In this way, treatment groups only differ in the application of the pre-test, and similarly for control groups.[6] Then, we can estimate the effect of the pre-test on the post-test scores under two scenarios: When the treatment is present and when it is absent. Under the first scenario we compare the mean post-test scores of groups T_1 and T_2; if the mean score of T_1 is significantly larger than that of T_2 then we may ascribe this difference to the pre-test. Similarly, we compare the scores of C_1 and C_2; if there is no difference in this scenario, but there is a difference in the previous scenario, then we can claim that the pre-test affected the post-test not in a direct way but via the treatment (Ariel et al., 2021); for example, students became aware or sensitive of the type of information they must seek from the tutor LLM and thus exploited this sensitization to intentionally perform better on the post-test. Had there been a difference in scores between C_1 and C_2 in a similar magnitude and same direction as that between T_1 and T_2 then we could have claimed that the pre-test affected directly the post-test (possibly via a mechanism from the memorization bias, or due to test practice). This design requires twice as much participants as the basic design (around 200 is recommended (Navarro and Siegel, 2018)); but when the target population is abundant, this is the suggested RCT design to carry out.

[6] We note that assignment of participants to the four groups must be done using a random allocation method where any participant has the same chance to be assigned to any group.

7 Statistical Analysis of RCT Data

To estimate the effect of the treatment we can resort to two simple and popular statistical methods: The difference of means between treatment and control groups or an Analysis of Covariance (ANCOVA). We elaborate on these methods by exemplifying them with the basic RCT design (our account can be easily expanded to more groups).

7.1 The Simplest Method to Estimate the Treatment Effect: Difference of Means

In the basic RCT design, the only difference between treatment and control groups is the treatment effect (Murnane and Willett, 2010);[7] thus, by comparing the mean post-test scores of the two groups we estimate the magnitude of the treatment effect (similarly to how the mean accuracy score of a new NLP model is compared to that from a baseline model). For example, if students in our tutor LLM group obtained a mean post-test score of 8.4 points while the control group mean score was 6.5 points, then the treatment effect was 1.9 points. Alternatively, if we want to estimate the treatment's effect as the mean difference in progress from pre-test to post-test, then we can compare the average change of scores from pre-test to post-test between our two groups (Ariel et al., 2021); i.e., for each group we obtain the change in scores (post-test score − pre-test score) averaged across participants, and compare the two groups' means. (For other details see Section A.5.)

7.2 ANCOVA: Quantifying the Effects of the Treatment and Confounders

ANCOVA is a multivariate linear regression analysis where the dependent variable (post-test scores or change in scores from pre-test to post-test) is determined by a linear combination of independent variables, namely the treatment variable and covariates (i.e. confounders[8]) (Wang et al., 2019); co-

[7] Assuming both an optimal random allocation of participants to both groups and no placebo effect in the treatment group.

[8] Covariates is one name that confounders receive in statistics.

variates such as gender, age, socioeconomic status, etc. may have an effect on the post-test scores, specially when the sample size is small, since due to chance they may not be equally distributed across groups (Torgerson and Torgerson, 2008); moreover, including covariates in an ANCOVA model not only allows us to quantify their effect on the post-test scores, but also allows us to isolate them from the treatment effect (Read et al., 2013); furthermore, we can also estimate the effect of the pre-test on the post-test scores. This specification helps to improve the precision of the treatment effect estimation (Egbewale et al., 2014).

Equation 1 shows ANCOVA's model:

$$y = \mu + \beta_1(x_1) + \beta_2(x_2) + ... + \beta_T(x_T) + \epsilon \quad (1)$$

Variable y represents the dependent variable; term μ is the intercept of the linear model representing the expected score of a control participant when all covariates are set to zero. Coefficients β_i ($i \in \{1, 2, ..., n\}$) are parameters to be estimated and represent the effects of covariates x_i on the dependent variable; if a covariate is continuous, its coefficient represents the mean number of points that will be added to the dependent variable for each unit that the covariate increases; x_T is a binary indicator variable where $x_T = 1$ represents membership to the treatment group and $x_T = 0$ to the control group; thus, β_T is the estimated treatment effect which we interpret as the average number of points that a treatment participant will gain due to receiving the treatment. Interpretation of coefficients of categorical covariates is similar to that of the treatment variable. Finally, ϵ represents the error term.

Let us propose in Equation 2 an ANCOVA model obtained from our hypothetical RCT study testing the efficacy of our diagnostics tutor LLM where the dependent variable represents post-test scores.

$$score = 4 + 0.8(test) + 0.5(SES) + 2(LLM) \quad (2)$$

We include the covariate of socioeconomic status (SES) of students as an indicator variable where $SES = 0$ represents a low-SES student and $SES = 1$ a high-SES student; also, we model the effect of the pre-test score as a continuous variable; the treatment variable is a binary indicator where $LLM = 1$ indicates membership to the treatment group and $LLM = 0$ membership to the control group.

We interpret the model in Equation 2 as follows: The expected post-test score of a control student from low-SES with 0 points on the pre-test is 4 points. The effect of the treatment has a magnitude of 2 points, after adjusting for SES and pre-test; this means that students who have the support of the LLM ($LLM = 1$), compared to those who do not, score on average 2 points more on the post-test. Moreover, the pre-test has an effect on students possibly due to the sensitization or memorization bias depending on the exact form of the pre-test (Section 4.1); we interpret its coefficient as adding, on average, 0.8 points to the post-test score for each point obtained in the pre-test (after adjusting for SES and treatment group). Furthermore, after adjusting for pre-test and treatment group, high-SES students tend to obtain, on average, 0.5 points more on the post-test than low-SES students. (For details of ANCOVA see Section A.6.)

8 Ethical Concerns in RCTs for LLMs

LLMs are a recent technology. The study of their interaction with different user types across clinical domains has just started and significant further research is necessary to thoroughly understand the LLMs' effects; thereby, unexpected results may be obtained in any RCT study. Consequently, we recommend to first carry out RCTs on problems or conditions where the risks of adverse results are low, and avoid critical scenarios such as adolescents with severe depression since any negative side-effect could worsen their initial condition, such as becoming psychologically distressed or even suicidal. Critically, in the event where the LLM treatment were found to be adverse, it would be imperative to both assess the outcomes to decide whether to stop or continue the study and report these negative outcomes (Anderson et al., 2022) to allow the scientific community to ponder the adversities.

Moreover, RCTs are not without any ethical concerns. One common issue is the so-called *compensatory equalization* problem which refers to the unequal situation created under the basic RCT design (Section 6.1) or the placebo RCT design (Section 6.2), where only one group receives a potential benefit (a treatment), while the other group receives no-treatment and thereby no potential so-

lution to their underlying condition or problem (Creswell and Guetterman, 2018). This situation can be solved using either of the first two solutions proposed to control the *resentful demoralization* bias (Section 5.3), namely to give a traditional, helpful treatment to the control group (such as a human teacher or counselor if the novel treatment is an LLM), or to offer the novel treatment to the control (or placebo) group after the experiment has finished.

Another ethical concern in RCTs is the effect of the novel treatment. The usual hypothesis is that the novel treatment will provide a positive and helpful effect on the participants; however, in reality, the treatment effect may be negative or may have adverse secondary effects (Bishop and Thompson, 2023). Hence, it is recommended to continuously observe the participants to spot any sign of negative or adverse reactions in the participants' behaviors, and terminate the study if those reactions are found.

Finally, we strongly emphasize that any RCT study, prior to its deployment, must be reviewed and approved by an ethics committee from a university or an institutional board (Creswell and Guetterman, 2018), and consent must be obtained from all the parties involved in the study including the participants and any parents or legal guardians if required (Moher et al., 2010). Ultimately, participants have the right to drop out of the study at any time, and any action to force the participant to not do so must not be taken. We recommend (Moher et al., 2010) for a next guide on RCTs.

9 Conclusions

Evaluating the efficacy or usefulness of a new medical treatment, educational intervention, or therapy program in Medicine, Psychiatry, among other clinical fields, is done via Randomized Controlled Trials. This strict evaluation has received the nomenclature of evidence-based evaluation in those fields given that RCTs are the experimental method considered pertinent to provide rigorous scientific evidence for the efficacy of a target treatment. As exposed by Reiter (2025), the field of NLP is practically devoid of this type of evaluation for NLP applications. Crucially, we consider RCTs are necessary to assess recent NLP applications such as general-purpose LLMs–and more critically, biomedical or clinical support LLMs–which are no longer laboratory objects but everyday tools, and have shown both positive and negative impacts by helping users with routine tasks or by suggesting the ingestion of chemicals with potentially toxic characteristics, for example. As such, other fields, prominently Medicine, have recently started to evaluate LLMs across tasks and capabilities to provide scientific evidence of their benefit (or not) to specific populations. The NLP field is to take the next step in LLMs evaluations via RCTs as it is the designer and builder of these NLP tools. In this primer paper, we presented the basic principles and statistical methods required for designing Randomized Controlled Trials targeted to assess LLMs' efficacy for real-world cases, helping bio-NLP researchers to take the first step to conduct RCTs.

Limitations

As a primer paper, this work mainly shows content and scope limitations. It lacks advanced methods used in specific problems faced by RCTs such as better random assignment methods of participants to groups or a more comprehensive treatment of the placebo bias, more elaborate or complex RCT designs, recently discovered biases, a wider and more in-depth treatment of ethical concerns in RCTs, or state-of-the-art statistical methods for estimating treatment effects. Nonetheless, we note that advanced methods for each of these issues may be problem- or field-oriented; therefore, by providing a foundational treatment of the design of RCTs, this paper serves as the starting point to discovering advanced methods in specialized literature and to proposing methods tailored for the biomedical NLP field.

Ethical Considerations

We do not claim that this paper is a comprehensive, self-contained, and ultimate guide for designing and conducting RCTs. This is a primer paper that aims to introduce biomedical NLP researchers to the foundational principles of RCTs in order to advance to other specialized guides in other fields, to start the intellectual task of designing and proposing RCTs for evaluating LLMs, and to foster interdisciplinary works with researchers in clinical disciplines. Therefore, we advise the use of this primer paper as an introductory guide–but not the sole resource–for biomedical NLP researchers to start their study in designing RCTs for assessing LLMs while subsequently consulting more advanced texts and research papers.

Acknowledgments

We thank the reviewers for their valuable comments. This work was supported by Beijing Natural Science Foundation (IS25068).

References

John G. Adair, Donald Sharpe, and Cam Loi Huynh. 1990. The placebo control group: An analysis of its effectiveness in educational research. *The Journal of Experimental Education*, 59(1):67–86.

Alan Agresti and Maria Kateri. 2022. *Foundations of Statistics for Data Scientists: With R and Python*. CRC Press.

J. Michael Anderson, Conner Howard, Jessica Hardin, Cole R. Phelps, Chad Hanson, Reece M. Anderson, Matt Vassar, and Jake X. Checketts. 2022. Harms-related data are poorly reported among randomized controlled trials underpinning the american academy of orthopaedic surgeons clinical practice guideline recommendations for rotator cuff injuries. *Journal of Shoulder and Elbow Surgery*, 31(12).

Barak Ariel, Matthew Bland, and Alex Sutherland. 2021. *Experimental Designs*. SAGE Publications Ltd.

American Psychological Association. 2025. Artificial intelligence in mental health care. Last accessed: 2026-05-17. https://www.apa.org/practice/artificial-intelligence-mental-health-care.

Dawn M. Aycock, Matthew J. Hayat, Ashley Helvig, Sandra B. Dunbar, and Patricia C. Clark. 2018. Essential considerations in developing attention control groups in behavioral research. *Research in Nursing & Health*, 41(3):320–328.

Bernard C. Beins and Maureen A. McCarthy. 2018. *Research Methods and Statistics in Psychology*, 2 edition. Cambridge University Press.

Dorothy V. M. Bishop and Paul A. Thompson. 2023. *Evaluating What Works: An Intuitive Guide to Intervention Research for Practitioners*. Chapman and Hall/CRC.

Charlotte R. Blease. 2018. Psychotherapy and placebos: Manifesto for conceptual clarity. *Frontiers in Psychiatry*, 9.

Kelly M. Boone, Mark A. Klebanoff, Lynette K. Rogers, Joseph Rausch, Daniel L. Coury, and Sarah A. Keim. 2022. Effects of omega-3-6-9 fatty acid supplementation on behavior and sleep in preterm toddlers with autism symptomatology: Secondary analysis of a randomized clinical trial. *Early Human Development*, 169:105588.

Sophie Carruthers, Andrew Pickles, Tony Charman, Helen McConachie, Ann Le Couteur, Vicky Slonims, Patricia Howlin, Rachel Collum, Erica Salomone, Hannah Tobin, Isobel Gammer, Jessica Maxwell, Catherine Aldred, Jeremy Parr, Kathy Leadbitter, and Jonathan Green. 2024. Mediation of 6-year mid-childhood follow-up outcomes after pre-school social communication (pact) therapy for autistic children: randomised controlled trial. *Journal of Child Psychology and Psychiatry*, 65(2):233–244.

Wim Ceelen and Kjetil Soreide. 2023. Randomized controlled trials and alternative study designs in surgical oncology. *European Journal of Surgical Oncology*, 49(8):1331–1340.

Chen Chen, Kok Tai Lam, Ka Man Yip, Hung Kwan So, Terry Yat Sang Lum, Ian Chi Kei Wong, Jason C Yam, Celine Sze Ling Chui, and Patrick Ip. 2025. Comparison of an ai chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: Pilot randomized controlled trial. *JMIR Hum Factors*, 12:e65785.

John Creswell and Timothy Guetterman. 2018. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*, 6 edition. Pearson.

Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the New Statistics*, 1 edition. Routledge.

Bolaji E. Egbewale, Martyn Lewis, and Julius Sim. 2014. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Medical Research Methodology*, 14.

Audrey Eichenberger, Stephen Thielke, and Adam Van Buskirk. 2025. A case of bromism influenced by use of artificial intelligence. *Annals of Internal Medicine: Clinical Cases*, 4(8):e241260.

Paul Enck and Stephan Zipfel. 2019. Placebo effects in psychotherapy: A framework. *Frontiers in Psychiatry*, Volume 10 - 2019.

Xinyu Feng, Lidan Tian, Grace W K Ho, Janelle Yorke, and Vivian Hui. 2025. The effectiveness of ai chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: Systematic review and meta-analysis. *J Med Internet Res*, 27:e79850.

Lawrence M. Friedman, Curt D. Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. 2015. *Fundamentals of Clinical Trials*, 5 edition. Springer Cham.

Jens Gaab, Joe Kossowsky, Ulrike Ehlert, and Cosima Locher. 2019. Effects and components of placebos with a psychological treatment rationale – three randomized-controlled studies. *Scientific Reports*, 9(1):1421–1428.

Wenyi Gan, Jianfeng Ouyang, Hua Li, Zhaowen Xue, Yiming Zhang, Qiu Dong, Jiadong Huang, Xiaofei Zheng, and Yiyi Zhang. 2024. Integrating chatgpt in orthopedic education for medical undergraduates:

Randomized controlled trial. *Journal of Medical Internet Research*, 26:e57037.

AlMothana Gasaymeh, Asma'a Abu Qbeita, Reham AlMohtadi, and Mohammad Beirat. 2025. Exploring education students' use of chatgpt for academic and personal purposes: insights from a developing country context. *Frontiers in Education*, Volume 10 - 2025.

Jonathan Green, Kathy Leadbitter, Ceri Ellis, Lauren Taylor, Heather L Moore, Sophie Carruthers, Kirsty James, Carol Taylor, Matea Balabanovska, Sophie Langhorne, Catherine Aldred, Vicky Slonims, Victoria Grahame, Jeremy Parr, Neil Humphrey, Patricia Howlin, Helen McConachie, Ann Le Couteur, Tony Charman, and 2 others. 2022. Combined social communication therapy at home and in education for young autistic children in england (pact-g): a parallel, single-blind, randomised controlled trial. *The Lancet Psychiatry*, 9(4):307–320.

Allan Hackshaw. 2009. *A Concise Guide to Clinical Trials*, 1 edition. John Wiley & Sons, Ltd.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Anthony M H Ho, Rachel Phelan, Glenio B Mizubuti, John A C Murdoch, Sarah Wickett, Adrienne K Ho, Vidur Shyam, and Ian Gilron. 2018. Bias in before-after studies: Narrative overview for anesthesiologists. *Anesthesia and Analgesia*, 126(5):1755–1762.

Munier Hossain. 2021. *Making Sense of Medical Statistics: A Bite Sized Visual Guide*. Cambridge University Press.

Siyu Huang, Chang Wen, Xueying Bai, Sihong Li, Shuining Wang, Xiaoxuan Wang, and Dong Yang. 2025. Exploring the application capability of chatgpt as an instructor in skills education for dental medical students: Randomized controlled trial. *Journal of Medical Internet Research*, 27:e68538.

Kirsten Hyldgaard. 2020. The placebo effect in education? evidence based educational practice and the psychoanalytic concept of transference. *Journal of the International Society for Teacher Education*, 24(2):7–18.

Michael Kjær Jacobsen, Andreas Killerich Andresen, Annette Bennedsgaard Jespersen, Christian Støttrup, Leah Y. Carreon, Søren Overgaard, and Mikkel Ø Andersen. 2020. Randomized double blind clinical trial of abm/p-15 versus allograft in noninstrumented lumbar fusion surgery. *The Spine Journal*, 20(5):677–684.

Alejandro R. Jadad and Murray W. Enkin. 2007. *Randomized Controlled Trials: Questions, Answers, and Musings*, 2 edition. John Wiley & Sons, Ltd.

Caroline Barkholt Kamp, Sebastian Simonsen, and Sophie Juul. 2026. External validity of randomised clinical trials in psychiatry: borderline personality disorder as an example. *The British Journal of Psychiatry*, page 1–3.

Barbara Kingsley and Julia Robertson. 2020. *Your A to Z of Research Methods and Statistics in Psychology Made Simple*. Oxford University Press.

Irving Kirsch, Bruce Wampold, and John M. Kelley. 2016. Controlling for the placebo effect in psychotherapy: Noble quest or tilting at windmills? *Psychology of Consciousness: Theory, Research, and Practice*, 2(3):121–131.

Quan Li. 2018. *Using R for Data Analysis in Social Sciences: A Research Project-Oriented Approach*. Oxford University Press.

Peter Lilliengren. 2023. A comprehensive overview of randomized controlled trials of psychodynamic psychotherapies. *Psychoanalytic Psychotherapy*, 37(2):117–140.

Emma Marsden and Carole J. Torgerson. 2012. Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, 38(5):583–616.

S.E. Maxwell, H.D. Delaney, and K. Kelley. 2017. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 3 edition. Routledge.

Ryan K. McBain, Robert Bozick, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Joshua Breslau, Bradley D. Stein, Ateev Mehrotra, Lori Uscher Pines, Jonathan Cantor, and Hao Yu. 2025. Use of generative ai for mental health advice among us adolescents and young adults. *JAMA Network Open*, 8(11):e2542281–e2542281.

David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, P J Devereaux, Diana Elbourne, Matthias Egger, and Douglas G Altman. 2010. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340.

Sarah A. Moore, Djordje G. Jakovljevic, Gary A. Ford, Lynn Rochester, and Michael I. Trenell. 2016. Exercise induces peripheral muscle but not cardiac adaptations after stroke: A randomized controlled pilot trial. *Archives of Physical Medicine and Rehabilitation*, 97(4):596–603.

Anne M Moseley and Marina B Pinheiro. 2022. Research note: Evaluating risk of bias in randomised controlled trials. *Journal of Physiotherapy*, 68(2):148–150.

Richard Murnane and John Willett. 2010. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

Mario Navarro and Jason Siegel. 2018. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, chapter Solomon Four-Group Design. SAGE Publications, Inc.

Steven Piantadosi. 2017. *Clinical Trials: A Methodologic Perspective*, 3 edition. Wiley.

L. Popp and S. Schneider. 2015. Attention placebo control in randomized controlled trials of psychosocial interventions: theory and practice. *Trials*, 16(1):150–152.

Kendra L. Read, Philip C. Kendall, Mathew M. Carper, and Joseph R. Rausch. 2013. Statistical methods for use in the analysis of randomized clinical trials utilizing a pretreatment, posttreatment, follow-up (ppf) paradigm. In *The Oxford Handbook of Research Strategies for Clinical Psychology*. Oxford University Press.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

S. Sanders. 2019. A brief guide to selecting and using pre-post assessments. Technical report, American Institutes for Research, The National Technical Assistance Center for the Education of Neglected or Delinquent Children and Youth, Washington, DC.

Ekaterina Schneider, Cristóbal Hernández, Robert Brock, Monika Eckstein, Guy Bodenmann, Markus Heinrichs, Ulrike Ehlert, Severin Läuchli, and Beate Ditzen. 2025. Intranasal oxytocin and physical intimacy for dermatological wound healing and neuroendocrine stress: A randomized clinical trial. *JAMA Psychiatry*, 83(2):118–127.

Gary Smith. 2015. *Essential Statistics, Regression, and Econometrics*, 2 edition. Academic Press.

Richard L. Solomon. 1949. An extension of control group design. *Psychological Bulletin*, 46(2):137–150.

PM Spieth, AS Kubasch, AI Penzlin, BM Illigens, K Barlinn, and T Siepmann. 2016. Randomized controlled trials – a matter of design. *Neuropsychiatric Disease and Treatment*, 12:1341–1349.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

David J. Torgerson and Carole J. Torgerson. 2008. *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*. Palgrave Macmillan London.

Rebecca Tutino, Elizabeth Schofield, Rebecca M. Saracino, Leah Walsh, Emma Straus, and Christian J. Nelson. 2024. A survey of statistical methods utilized for analysis of randomized controlled trials of behavioral interventions. *Palliative and Supportive Care*, 22(2):221–225.

Gerard J.P. Van Breukelen. 2006. Ancova versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59(9):920–925.

Andrew J. Vickers. 2005. Parametric versus nonparametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, 5(35).

Fei Wan. 2021. Statistical analysis of two arm randomized pre-post designs with one post-treatment measurement. *BMC Medical Research Methodology*, 21(150).

Bingkai Wang, Elizabeth L. Ogburn, and Michael Rosenblum. 2019. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*, 75(4):1391–1400.

Jin Wang and Wenxiang Fan. 2025. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(621).

Herbert I. Weisberg. 2010. *Bias and Causation: Models and Judgment for Valid Comparisons*. Wiley.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Stephen G. West, Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, José Szapocznik, Martin Fishbein, Bruce Rapkin, Michael Clatts, and Patricia Dolan Mullen. 2008. Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98(8):1359–1366. PMID: 18556609.

Felix A Weuthen, Nelly Otte, Hanif Krabbe, Thomas Kraus, and Julia Krabbe. 2025. Comparison of chatgpt and internet research for clinical research and decision-making in occupational medicine: Randomized controlled trial. *JMIR Form Res*, 9:e63857.

Xiaonan Yu, Wilson W.S. Tam, Paul T.K. Wong, Tai Hing Lam, and Sunita M. Stewart. 2012. The patient health questionnaire-9 for measuring depressive symptoms among the general population in hong kong. *Comprehensive Psychiatry*, 53(1):95–102.

A Appendix

A.1 Further Biases in Before-and-After Studies

Hawthorne effect Firstly discovered in early work of optimization studies, this bias takes effect on the participants behavior by increasing their effort, workload, or productivity due to their awareness of taking part in a research program and being observed by a researcher (Torgerson and Torgerson, 2008). Thus, part of the results obtained are due to the artificially increased effort of the participants. For example, students in our hypothetical study will work harder than usual when receiving support from the tutor LLM, regardless of their inner motivation, just due to the pressure imposed by the study's monitoring apparatus.

Maturation bias When a treatment is applied over a sustained period of time (usually in the range of months), natural biological or psychological changes can happen to the study participants which may have an effect on the post-test scores (Ho et al., 2018); these changes go under the umbrella terms of *maturation effects* or *temporal effects* (Torgerson and Torgerson, 2008). For example, in our hypothetical study of a counseling LLM, a maturation effect that can bias the effect of the counseling LLM is the natural self-healing of the study participants over time (Bishop and Thompson, 2023); thus, any improvement in their well-being post-test scores could be simply due to this temporal effect rather than due to the support of the LLM.

Demand characteristics bias This type of bias is elicited from the study participants after learning that they will take part in a novel treatment with the goal of improving their skills or solving their underlying problem or condition (Beins and McCarthy, 2018). Participants actively over-engage in the study by increasing their motivation or eagerness which, in turn, may improve their focus and their performance on the study's tasks; for example, students in our hypothetical study may pay more attention to the LLM's outputs than they usually pay to their professors at class, may put more effort and time for out-of-study learning activities, or may ask for extra-help from peers or professors, and so on. This participants' quest to make the treatment effectively work can aid the participants in improving their post-test scores biasing the effect from the LLM alone.

Regression to the mean Some effects occurring in a study may happen only due to chance; these effects are termed *regression to the mean* (Marsden and Torgerson, 2012). For example, in our hypothetical study of the tutor LLM, there is some likelihood that due to chance some of the students will obtain a low pre-test score, and these students will naturally increase their score at post-test time due to their normal intellectual or cognitive capability; consequently, this improvement will be wrongly attributed to the LLM's help; in other words, students who, due to chance, scored particularly high or low at the pre-test will tend to score closer to the group's mean at the post-test (Marsden and Torgerson, 2012; Torgerson and Torgerson, 2008) biasing the effect of the LLM.

History bias There are so-called history events occurring outside the laboratory where the treatment is applied which are not controllable by the researcher. This bias occurs when these events, happening in parallel to the study, affect either the status of the treatment or the participants engagement in the target tasks impacting on the results obtained (Marsden and Torgerson, 2012; Ho et al., 2018). For example, adverse weather leading to temporarily closing the laboratory can directly affect the participants engagement in the study possibly leading to a distortion of the post-test scores and also of the treatment effect.

Experimenter bias In this case, it is the researchers executing the study the ones biasing the participants outcomes (Torgerson and Torgerson, 2008); for example, if researchers firmly believe that the support counseling LLM will be useful for diminishing the symptoms of adolescents suffering anxiety, then these researchers may unconsciously work towards fulfilling their believe by modifying the well-being questionnaires or giving additional support to the participants not planned before the study with the aim of generating outcomes that favor the counseling LLM.

Measurement bias This bias can be present when taking measurements, scoring a test, or when using an incorrect instrument to elaborate the tests (Creswell and Guetterman, 2018), depending on the research problem. For example, if a test (pre-test or post-test) is not correctly designed to measure the intended skill, ability, behavior, or mental/physical state of the participants, then the test will result in an invalid outcome giving a biased result; follow-

ing our hypothetical example of a counseling LLM, if the well-being questionnaire contains ambiguous questions related to the symptoms of the participant, the given answers may not be valid and so may biased the treatment effect.

A.2 Internal Validity of RCTs

The golden goal in any scientific experiment is to obtain valid results (Weisberg, 2010). When we test a new treatment to see its efficacy, we aim to obtain an unbiased relationship between the treatment and the results (post-test scores); i.e. we seek to answer the question: Is the new treatment the real cause of the results obtained? For example, is a counseling LLM really helpful to alleviate the symptoms of a group of adolescents suffering from depression? When answering this question, we expect that the effect obtained on the participants is a true and direct cause from the LLM and not an effect from a bias or a confounder. Hence, to answer this type of question we need an experimental design embodying a necessary characteristic: Internal validity (Beins and McCarthy, 2018), which refers to the capacity of providing valid conclusions, i.e. to neutralize the effects of spurious factors, such as those shown in Section 4.1, that can distort the true treatment effect. Arguably, RCTs' experimental design is one of the study designs with the highest internal validity (West et al., 2008) which is the main reason why RCTs are considered the gold-standard for evaluating new treatments in the fields of Medicine (Torgerson and Torgerson, 2008) and Psychiatry (Spieth et al., 2016), among other fields; and we hope RCTs will become the gold-standard for evaluating LLMs in the NLP field as well.

A.3 Matched Randomization

Using simple randomization we risk that due to chance we may end up with unbalanced groups in one (or more) confounder(s), especially when the sample size is small (30 participants or less (Torgerson and Torgerson, 2008)), which can lead to invalid results. For example, in our tutor LLM study, we hypothesize that socioeconomic status has a significant impact on post-test scores since students in a high socioeconomic status may have access to extra learning resources, such as private lessons or specialized books. Hence, not thoroughly controlling this confounder may give us an inflated (or diluted) treatment effect if students with a high socioeconomic status are mostly allocated to the treatment (or control) group. The method of matched randomization consists of pairing participants with the same level of a confounder and randomly assigning each participant from each pair to treatment and control groups to balance them (Creswell and Guetterman, 2018); this can be extended to more confounders. For example, if we hypothesize that besides socioeconomic status, gender is another confounder, then we create pairs of students across combined levels of these two characteristics: (high, low) $\times$ (female, male); for instance, we match two male, high-socioeconomic-status students, then toss a coin and randomly assign them to treatment and control groups; and similarly for all combinations of levels. Due to its simplicity, matched randomization is a popular method; however, when the number of confounders increases, matching can become restrictive since it may be difficult to find an exact match for a given participant across all levels. In this case, it is recommended to increase the sample size to at least 100 participants and instead implement simple randomization(Torgerson and Torgerson, 2008). Alternatively, we can use a statistical method to control confounders that only requires simple random allocation, as shown in Section 7.2.

A.4 Attrition Bias

There are events that are unforeseeable by the researcher, such as participants dropping out of a study. The primary way how attrition invalidates a study is by inducing selection bias, regardless of participants dropping out from the treatment, control, or both groups, since it breaks the balance of other types of bias and confounders imposed by the random assignment of these participants to the groups (Torgerson and Torgerson, 2008). Attrition occurring in both groups does not guarantee to keep the balance between them since the participants dropping out of one group can have different characteristics from those dropping out of the other group. Moreover, when attrition occurs in one group it is not recommended to remove participants from the other group which have similar observable characteristics since they may not have similar unobservable characteristics. For example, let us suppose that the youngest adolescents in our hypothetical study dropped out of the treatment group because they believed that the counseling LLM was not helpful for reducing their anxiety symptoms; removing the youngest participants from the control group to balance the groups again risks inducing selection bias if dropouts from the treatment group

differ from the youngest control participants in the level of socioeconomic status, or in any other characteristic. To reduce the impact of attrition, it is suggested to implement the policy of Intention-to-Treat (ITT) (Bishop and Thompson, 2023), which advises researchers to keep track of dropouts, with their consent, to obtain their post-test scores since including them in the analysis of results will keep the balance between treatment and control groups imposed by the random allocation. For a deep treatise on ITT we recommend (Torgerson and Torgerson, 2008).

A.5 Details of Estimating the Treatment Effect Via Difference of Means

The estimated treatment effect obtained via the difference of means is free from the effects of bias and confounders; however, it may still have been affected by random variation, i.e. by chance, (Hackshaw, 2009) present in several factors. To account for this randomness we need a statistical test; in this case a common test is the t-test (Murnane and Willett, 2010), which we use to obtain a statistic called the t-statistic (the treatment effect normalized by the variability between the groups), which in turn we use to estimate the p-value which we could interpret as the likelihood that the treatment effect could have been obtained by chance (Hackshaw, 2009). The logic behind the estimation of a p-value is that we compare our t-statistic against a distribution (a t-distribution) of differences between the two group means built under the assumption that, in fact, there is no difference between the two means (i.e. the treatment effect is zero); if our p-value is less than an established threshold then it means that it is unlikely that we would observe such a difference in means under this distribution, so we can claim that this effect seems not to be due to chance (Hossain, 2021). Alternatively, we can estimate a *confidence interval* which is a range of values in which there is certain likelihood that the population treatment effect falls into; if zero is not inside this range, then we can claim that the treatment effect seems to not be due to chance as well (Cumming and Calin-Jageman, 2016). We suggest (Hackshaw, 2009; Hossain, 2021) for an elaboration of this statistical method and (Agresti and Kateri, 2022; Li, 2018) to estimate the t-statistic and its p-value with a statistical software as is commonly done in the literature.

While this is the easiest method to estimate the treatment effect, it has a drawback if our sample size of participants was small: It cannot measure the effects of confounders that may be lurking in the results (due to chance imbalance) which alter the treatment effect. So, unless we have a big sample size of participants and we carefully apply random allocation, and even so, the recommended statistical method for analyzing RCT data is ANCOVA (Van Breukelen, 2006; Vickers, 2005).

A.6 Details of ANCOVA

ANCOVA is one of the predominant statistical methods to analyze RCT data (Tutino et al., 2024) due to 1) its simplicity (Wan, 2021), 2) its flexibility to account for both types of variables (continuous and categorical) which allows us to estimate the treatment effect (a categorical variable) and the covariates' effects (categorical or continuous variables) (Maxwell et al., 2017), and 3) we can also estimate the effect of the pre-test on the post-test scores which results in a more precise estimate of the treatment effect (Egbewale et al., 2014).

The use of ANCOVA assumes some properties such as the linear relationship between dependent and independent variables, the dependent variable being normally distributed, or the variability of the groups being homogeneous; however, ANCOVA has been shown to be robust to departures from such assumptions (Wang et al., 2019).

To evaluate the fit of an ANCOVA model, we can use the mean-squared error: The averaged squared difference between predicted and true post-test scores, or the R^2 metric which estimates how much of the variability in the post-test scores is explained by the treatment and covariates, i.e. it measures the proportional reduction in the difference between predicted and real scores (Agresti and Kateri, 2022). For a deeper treatise on ANCOVA, its evaluation, and its extension to more groups, we suggest (Wang et al., 2019; Maxwell et al., 2017), and (Agresti and Kateri, 2022; Li, 2018) for implementing ANCOVA with a statistical software.